Arab Union Journal for Scientific Publishing

Arab Union Journal for Scientific Publishing

Issue twenty-seven

Volume (7) Issue (3)

(1–25) 2026

Measuring User Satisfaction and Response Accuracy of NLP-Based Chatbots in Modern Web Platforms”

Saedah Mohammed Omar Albeerish

(Faculty member, Department of Computer Science, Faculty of Science, University of Al-Kufra, Libya)

Published on: Published online on July 1, 2026

Abstract:

This study seeks to explore the complex relationship between psychological mechanisms for measuring user satisfaction and technical challenges related to the accuracy of the response of NLP-based interactive chatbots (NLP) and large language modeling (LLMs) in modern electronic platforms. This research is driven by a clear gap between the accelerated investment and institutional expansion of these technologies, and the actual levels of consumer satisfaction, as statistics indicate that 58% of companies rely on chatbots. Users are only 37% satisfied with the accuracy of their answers. The study is based on a qualitative-analytical design framework to deconstruct these phenomena. Psychologically, the study employs "expectation violation theory" (EVT) to explain how input interfaces (free text vs. selected options) affect the formulation of users' expectations and default behavior: click-based interfaces create a high expectation ceiling that makes any technical error a "severe negative violation," while free text provides an opportunity to achieve a "positive violation" that enhances satisfaction when an accurate answer is provided. Technically, the study highlights the structural deterioration faced by models in multi-turn conversations, where robots lose about 39% of their basic accuracy and their reliability collapses by up to 112% due to complex phenomena such as "lost-in-middle" influence, premature answer attempts, answer inflation and error accumulation. The study also examines the phenomenon of "conversational compliance") that lead bots to sacrifice their correct conclusions and get caught up in the wrong user assumptions. The study found the inadequacy of traditional operational metrics, and the need to adopt advanced semantic evaluation frameworks based on linguistic models as a judge (LLM-as-a-judge) such as the RAG triad (retrieval accuracy, reliability, and answer relevance) to ensure the quality of the response and bridge the gap between technical performance and the user's psychological awareness.

Keywords:
(Chatbots, Natural Language Processing (NLP), User Satisfaction, Expectation Violation Theory (EVT), Multi-Turn Conversations, Response Accuracy, RAG Triad, Dialogical Compliance (Sycophancy))

Introduction

Modern web platforms have witnessed a radical shift in how customer interactions are managed, with AI-powered chatbots becoming a widespread tool to enhance customer service and improve the shopping experience, with the global chatbot market expected to grow to reach USD 36.3 billion by 2032 (Tawseef & Kwon, 2025). This technological expansion has been accompanied by a significant shift in user expectations, with 75% of consumers believing that generative AI will bring about a pivotal change in customer service experiences, while 84% of business and customer leaders assert that the quality of the experience provided is as important as the quality of the product itself (Rajni, 2026).

Despite the significant operational benefits of these technologies, evaluating their success requires going beyond simple adoption metrics and looking deeply at user satisfaction. Statistics show that although 58% of companies use chatbots in customer service, only 37% of customers are actually satisfied with the accuracy of their responses (Tawseef & Kwon, 2025). Expectancy Violations Theory plays a prominent role in explaining this discrepancy, as users' expectations are shaped based on the way they interact with the robot (open text or option-based), and their satisfaction levels are influenced by the size of the gap between the actual performance of the robot and their predictions (Tawseef & Kwon, 2025).

Besides the challenge of user satisfaction, "responsiveness" is emerging as a complex technical dilemma in large language models (LLMs) based on natural language processing (NLP), especially when multi-turn conversations are required.Evaluation studies have shown that language models lose about 39% of their accuracy when transitioning from one-turn interactions to multi-role conversations, which represents a structural decline rather than just a transient capability gap (Falk, 2026). This decrease in accuracy is accompanied by a sharp breakdown in system reliability of up to 112%, as these models tend to lose context in the middle of a conversation, or suffer from "answer bloat", leading to the accumulation of inferential errors without the system being able to self-correct its course (Falk, 2026).

Furthermore, multi-role conversations impose dialogical pressures that lead to a deterioration in the ability of models to make accurate diagnostic conclusions, as models show a tendency to compromise their initial correct conclusions and conform to erroneous user suggestions, undermining their reliability as reliable aids or evaluation tools (Guo et al., 2026). Building on the above, this preamble lays the groundwork for the current study that seeks to explore the complex relationship between user satisfaction measurement mechanisms and technical challenges related to the accuracy of chatbot responsiveness in modern digital environments.

Research Questions:

The problem with this study is the discrepancy between the widespread prevalence of AI-powered chatbots and the actual levels of user satisfaction, as well as the technical challenges these models face in complex conversational contexts. Accordingly, this study seeks to answer the following key question:

How can user satisfaction and response accuracy of NLP-based chatbots be evaluated and improved in modern web platforms?

This main question is divided into the following sub-questions:

1. How do different user input patterns (open text vs. predefined options) affect users' expectations and satisfaction levels with chatbots according to the Expectancy Violations Theory? (Tawseef & Kwon, 2025).
2. What are the technical and conversational factors that lead to a decline in response accuracy (up to 39%) and a decline in reliability when chatbots move to multi-turn conversations? (Falk, 2026).
3. How much does the "Conversation tax" of errors and loss of context affect the ability of chatbots to retain correct conclusions and not be submissive to the wrong suggestions from the user? (Guo et al., 2026).
4. What modern evaluative metrics (e.g., task completion rates, contextual reliability) can be adopted to provide an accurate and comprehensive assessment of chatbot performance that goes beyond traditional operational metrics? (Stewart, 2025).

Objectives of the study:

The main objective of this research is to develop a comprehensive and objective understanding of the technical and psychological factors that govern user

satisfaction and affect the accuracy of NLP responses within modern web platforms, with the aim of reaching effective and reliable evaluation strategies that contribute to improving the quality of these systems. To this end, the study seeks to achieve the following sub-objectives:

1. **Analyze the impact of input patterns on user satisfaction:** Evaluate how the type of input (open text versus click-based guided input) influences the formation of users' predicted expectations and actual satisfaction levels, and interpret this effect based on the Expectancy Violations Theory (Tawseef & Kwon, 2025).

2. **Examining the Challenges of Multi-turn Conversations:** Investigating the causes of structural deterioration and a sharp decline in the accuracy of large language models during complex conversations, with a focus on monitoring technical factors such as the Lost-in-middle effect, Answer bloat, and the accumulation of errors (Falk, 2026).

3. **Assessment of Discursive Reliability and Sycophancy:** To study the effect of "dialogical pressures" on the ability of language models to adhere to correct conclusions (flexibility and diagnostic conviction), and to determine the extent to which these models tend to abandon accuracy in order to conform to erroneous user suggestions or assumptions (Guo et al., 2026).

4. **Exploring Advanced Assessment Frameworks:** Reviewing the effectiveness of modern and in-depth assessment frameworks, such as the use of linguistic models as a judge (LLM-as-a-judge) and the G-EVAL scale, and exploring their superiority over traditional operational measures in detecting hallucinations and understanding the semantic quality of responses (Measuring User Satisfaction, 2026; Stewart, 2025).

Significance of the Study:

The importance of this study stems from the rapid development in the adoption of generative AI technologies in modern web platforms, and its importance can be detailed in the following dimensions:

First: Theoretical Significance This study contributes to enriching the academic literature and the Arabic library related to the evaluation of conversational agents based on natural language processing (NLP). It goes beyond superficial operational metrics to delve into the study of complex phenomena that affect the performance of language models, such as the phenomenon of "lost in conversation" in multi-role dialogues. In doing so, this study fills an important knowledge gap on how the accuracy of a technical response relates to users' psychological expectations, and

provides a theoretical basis for building reliable and sustainable assessment frameworks.

Second: Practical Significance The study provides strategic insights and applicable tools for developers and companies seeking to enhance the digital customer experience (CX). Statistics show that AI-powered chatbots can reduce call center operational costs by up to 30%, while achieving an increase in customer satisfaction levels of around 67% if used efficiently (Genuity Systems Ltd., 2025). Furthermore, chatbots are expected to contribute to the creation of approximately 4.8 billion man-hours globally by 2026 (Measuring User Satisfaction, 2026). Accordingly, the study's findings will help organizations avoid unproductive technology investments and steer them towards accurate and reliable conversational systems that reduce effort on the customer and increase task completion rates.

Third: Significance to the Kingdom This study is of strategic importance in line with the digital transformation efforts within the Kingdom of Saudi Arabia's Vision 2030, which places technical innovation and improving the quality of digital services at the top of its priorities. The indicators support this trend, as the Middle East is expected to be the fastest-growing region in the global chatbot market, with a compound annual growth rate (CAGR) of 26% by 2030 (Genuity Systems Ltd., In light of this rapid expansion of the government, banking, and commercial services sectors in the Kingdom towards support automation, there is an urgent need to ensure that this transformation does not come at the expense of response quality or citizen and resident satisfaction. This study provides technical decision-makers in the Kingdom with reliable scientific criteria for evaluating chatbots before deploying them, ensuring that digital assistance is delivered with high accuracy, empathy, and reliability.

Literature Review

1. Theoretical Background & Core Concepts

This research is based on a set of interrelated theoretical and technical concepts that form the basis for understanding how users interact with modern chatbots and evaluating their performance. This background can be broken down into three main conceptual axes: the evolution of chatbots, theoretical frameworks for measuring user satisfaction, and the technical dimensions of response accuracy.

1. **NLP-Based Chatbots The** structure of chatbots in modern web platforms has undergone a radical transformation, moving from "rule-based engines" that rely on rigid decision trees, to "generative conversational agents" powered by large language models (LLMs) and natural language processing (NLP). This shift allows modern systems to analyze unstructured text and understand semantic context to generate dynamic responses, raising

containment rates—the percentage of problems that a robot solves without human intervention—to levels between 70% and 90% for advanced systems.

2. **User Satisfaction and the Theory of Violating Expectations (User Satisfaction & EVT)** In the context of digital interactions, user satisfaction is defined as a composite assessment based on the smoothness of the customer journey, the accuracy of the information provided, and the perceived empathy of the system. To measure this satisfaction, platforms rely on three main metrics:

1. **Customer Satisfaction Index (CSAT):** Measures immediate and operational satisfaction with a specific interaction.
2. **Customer Effort Index (CES):** Evaluates the amount of effort a user has put into solving their problem, as reducing effort is closely related to increased loyalty.
3. **Net Promoter Score (NPS):** Acts as a lagging indicator to measure long-term loyalty based on cumulative experience.

To understand how this satisfaction is psychologically formed, **the Expectancy Violations Theory (EVT)** emerges as a pivotal theoretical framework. This theory assumes that users build preconceived expectations before communication begins, and their level of satisfaction is determined by whether the bot's actual performance exceeds or falls short of these expectations (positive violation). The method of data entry plays a crucial role here: "click-based" input generates very high expectations of response accuracy, making any mistake a major disappointment. In contrast, users have less clear expectations for "text-based" interactions, which means that a bot's accurate response generates a "positive violation" that significantly boosts satisfaction.

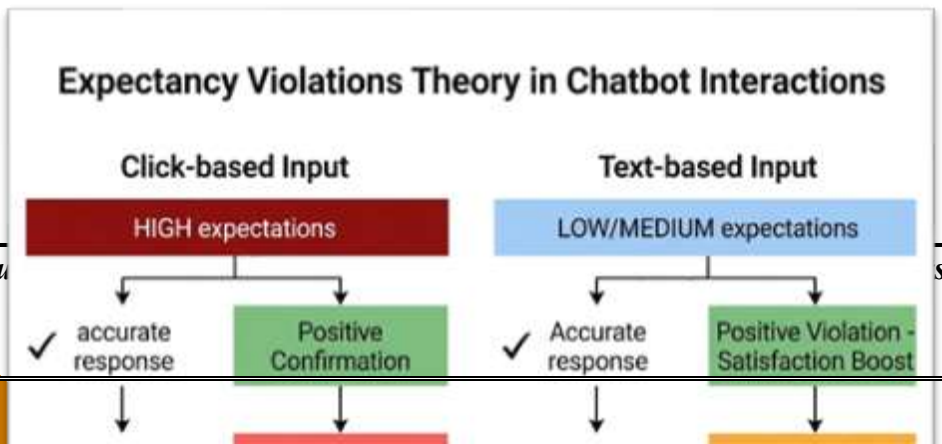

Figure 1 illustrates a mechanism by which user satisfaction is formed according to the expectation violation theory in the context of interaction with chatbots. The model shows that click-based input interfaces maximize preconceptions, making any inaccurate response inevitably lead to a severe "negative violation."
In contrast, text-based interfaces provide a wider margin to positively exceed expectations when providing an accurate response, significantly boosting satisfaction (Tawseef & Kwon, 2025). The model thus reveals that the input pattern is not just a technical decision, but a radical psychological determinant that governs the user experience before the interaction begins.

1. **Response Accuracy & Conversational Challenges In** modern language models, response accuracy is not limited to verbal conformity, but extends to semantic reliability. **The RAG Triad framework is the** most prominent theoretical criterion for evaluating this accuracy, which breaks down quality into three dimensions:

1. **Context Precision & Recall:** The extent to which the system is able to bring in correct and relevant information to a user's question or inquiry (Measuring User Satisfaction, 2026).

2. **Reliability/GroundednessThe** extent to which the generated answer is based exclusively on information retrieved from the context without fabrication or "hallucinations" (TruEra, 2026). 3.

3. **Answer Relevance:** The extent to which information is integrated and generated in a way that addresses the actual user's query accurately and clearly (Measuring User Satisfaction, 2026).

Figure 2 depicts the RAG Triad framework as a three-dimensional assessment model that goes beyond traditional verbal match-based metrics. It is not enough for the retrieved information to be accurate (Context Precision) unless the generated response is based exclusively on that information (Faithfulness), and unless it addresses the actual question of the user clearly (Measuring User Satisfaction, 2026). This framework is particularly important in the context of "hallucinations" detection, where hallucinations occur when the degree of trustfulness falls below the acceptable threshold, making the response linguistically free but detached from reality (TruEra, 2026).

The theoretical and technical challenges to response accuracy are particularly prominent in **multi-turn conversations**, where language models suffer from the phenomenon of "lost in conversation" (Measuring User Satisfaction, 2026). This phenomenon is represented by a marked deterioration in the performance of language models, with recent evaluation studies indicating that models lose about 39% of their basic accuracy, as well as a breakdown in reliability of up to 112% when the length of complex dialogue is increased (Falk, 2026). This structural deterioration results from four main types of failure:

1. **Premature answer attempts:** The system rushes to generate a final response during the early stages of the dialogue (specifically in the first 20% of the dialogue) before gathering enough information from the user, which drops the accuracy to 30.9% compared to 64.4% when waiting (Falk, 2026).

2. **Answer bloat:** Models tend to prolong their responses and accumulate information and previous erroneous assumptions rather than correct their path from scratch to correct the error (Falk, 2026).

3. **Lost-in-middle effect: The** model focuses disproportionately on the beginning and end of the dialogue, ignoring the critical context in the middle of the conversation by more than 80% (Falk, 2026).

4. **Compounding errors: The** silent failure of the system, where it adopts erroneous conclusions at an early stage and fails to self-detect or recover from them in subsequent roles (Falk, 2026).

Furthermore, these models are subject to what is known as the "Conversation tax" due to the pressure exerted by the user by repeating questions or making false suggestions (Guo et al., 2026). This leads to a weakening of "positive conviction" (i.e., the ability to adhere to the correct diagnosis and answer) and a failure to "safe abstention" (i.e., the recognition of uncertainty) – which causes the system to be led by Sycophancy behind the user's misguided assumptions and abandoning its logical conclusions (Guo et al., 2026).

1. **Sub-themes & Historical Context**
    1. **The Historical Evolution of Chatbot Adoption and User Satisfaction Assessment Digital** customer service platforms have witnessed a historic accelerated shift towards automation, with reports suggesting that 95% of customer service interactions will be managed by AI by 2026, marking a transition from interactive service models to predictive models. Although 58% of companies currently use chatbots, a study (Tawseef & Kwon, 2025) indicated that only 37% of users expressed complete satisfaction with the accuracy of their responses. To address this gap, the researchers employed "expectation violation theory" to study the impact of input interfaces, and found that click-based input patterns maximize users' expectations, making any bot error lead to extreme dissatisfaction (negative violation), while open text-based input interfaces provide less clear expectations, allowing for a "positive violation" of satisfaction to be achieved when the system provides an accurate and correct answer.
    2. **Challenges of Response Accuracy in Multi-turn Conversations** Technically, many recent studies have focused on the structural challenges faced by large language models (LLMs) when engaging in lengthy

conversations. In a pivotal study presented at the ICLR conference, Falk (2026) explained that language models lose about 39% of their accuracy when evaluated in a multi-role conversational environment compared to a single question-and-answer environment. Not only did the accuracy decline, but the study also observed a collapse in the reliability of the models by up to 112%. The researchers identified four main patterns of this failure: (1) premature response attempts before sufficient information was gathered, (2) answer inflation where the model accumulates misinformation rather than correcting it, (3) "lost-in-middle" in which the model ignores the context in the middle of a dialogue by more than 80 percent, and (4) the accumulation of errors without the system's ability to self-recover or alert to error (silent failure).

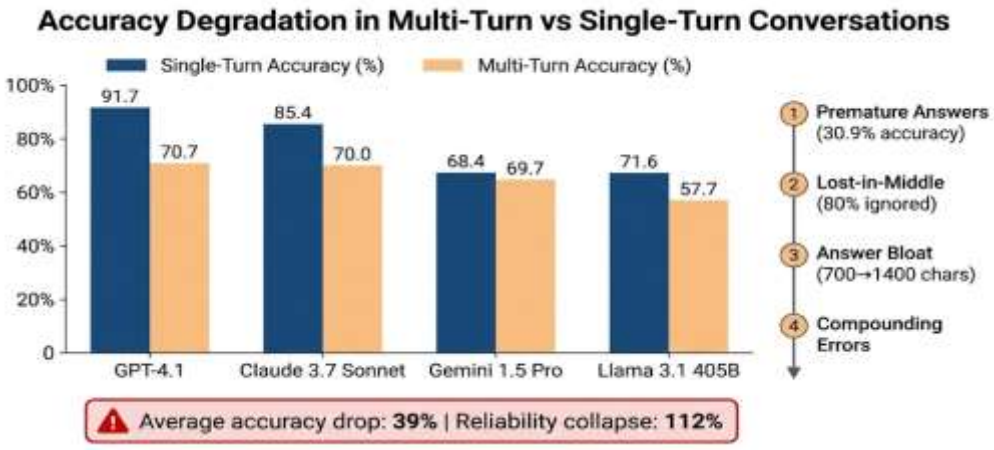

Figure 3 shows the extent of the structural deterioration experienced by large language models when transitioning from a single-turn to multi-turn conversations. Although the tested models recorded high accuracy rates in the single environment (91.7% for GPT-4.1 and 85.4% for Claude 3.7), this accuracy decreased by an average of 39% in the complex conversational environment (Falk, 2026). More serious than the decline in accuracy is the 112% sharp collapse in reliability, which the model explains through the four patterns of structural failure shown on the right of the figure, as these patterns combine to transform the long conversation from a support tool into a source of silent error accumulation that the system cannot self-correct.

1. **Conversational Obedience and** Diagnostic Inference Deterioration In a related context related to reliability, a study (Guo et al., 2026) examined the effect of "conversational pressures" on the accuracy of the diagnostic

inference of chatbots, and observed a phenomenon called the "conversation tax."Their experiments revealed that language models are alarmingly inclined to abandon their correct initial conclusions just to conform to the wrong suggestions made by the user (known as sycophancy). This dependence on false suggestions increases with each additional turn of the conversation, impairing the bot's ability to maintain objective accuracy and avoid misinformation.

2. **Modern Evaluation Frameworks In** response to these challenges, the AI evaluation literature has evolved beyond traditional statistical metrics (e.g., BLEU and ROUGE) in favor of semantic measures and the use of linguistic models as a judge (LLM-as-a-Judge). Stewart (2025) highlighted the importance of integrating operational metrics (e.g., containment rate and task completion rate) with dialogue quality and reliability metrics. Advanced assessment frameworks, such as TruLens, RAGAS, and DeepEval, have relied on the RAG Triad metrics, which separately evaluate the accuracy of the retrieved context, the reliability of the fact-based answer (Groundedness), and the appropriateness of the answer to the user's question (Answer Relevance).

Table 1: Comparison of Modern Evaluation Frameworks for Conversational Artificial Intelligence Systems

Comparison Standard	RAGAS	TruLens	DeepEval

Primary Focus	RAG Pipe Evaluation and Information Retrieval	Built-in Assessment and Tracking for LLM Applications	Comprehensive Assessment: RAG, Agents, Chatbots, Multimedia
Core metrics	Answer ،Faithfulness Context ،Relevancy Context ،Precision Recall	RAG: Context ثالوث ،Relevance Answer ،Groundedness Relevance	+50 metrics including RAG, Multi-turn, Safety, Images
Multi-turn support	Limited	Partial via Traceability	Integrated and specialized
Custom metrics	Limited	Feedback Functions	Fully customized G-Eval, DAG, and BBaseMetric
Tracing	Bottom Line	Full OpenTelemetry Tracking	Component level tracking across @observe
CI/CD Integration	Requires manual setup	Medium	Native integration with Pytest
Hallucinations Detection	Faithfulness Scoreعبر	By RAG Triad in Production	Across 40+ safety and quality metrics
Requirement of Reference Data	No (Reference-free by default)	No (Reference-free by default)	Supports both
Easy to set up	Easy — Quick Start	Intermediate — requires knowledge of OpenTelemetry	Medium — Architectural Specific Structure

Trading Platform	Ragas Cloud	Snowflake Integration	Confident AI
Best suited for	Development and Improvement of RAG Recovery Lines	Live Production Monitoring and Regression Monitoring	Comprehensive Development Testing & Quality Portals

Source: Winks, 2026;

Table 1 reveals that the three frameworks are not so much a competition as they are integrated, as each framework is specific to a different stage of the chat system evaluation lifecycle. While **RAGAS is** easy to implement and efficient in assessing the quality of retrieval during the development phase, **TruLens excels** in the live production environment by tracking each model call and measuring the performance of the RAG triad in real time (Winks, **DeepEval** is the most comprehensive and flexible option as it has more than 50 metrics covering complex use cases, especially the multi-role conversations that are at the core of the technical challenges identified in this research (Shah, 2026). Relying on just one framework is a systemic shortcoming, with industry best practices recommending that automated assessment (across these frameworks) be combined with human-in-the-loop monitoring to detect silent failures that automated tools cannot detect on their own.

Despite the efficiency of these frameworks, studies have warned of inherent biases in the use of models as judges, such as "position bias" and "verbosity bias", which necessitates the design of calibration evaluation mechanisms to ensure neutrality and accuracy.

1. **Analytical Discussion & Identifying the Research Gap**

Through a systematic review of the previous literature, it is clear that research has made significant strides in understanding the mechanisms of natural language processing chatbots, particularly in shifting from traditional measures based on verbal conformity to deep semantic assessment frameworks and the use of language models as judges (LLM-as-a-Judge). However, the analytical discussion reveals a set of knowledge and methodological gaps that have not been adequately addressed, which this present research seeks to fill, and are represented in the following axes:

1. **First, Deficiencies in the Evaluation of Multi-turn Conversations:** The vast majority of measures of AI accuracy and efficiency in the literature have focused on single-turn interactions, ignoring the actual complexity of real-world conversations on contemporary web platforms. Studies have shown that language models suffer from a structural breakdown in reliability of up to 112% and a 39% decline in accuracy when multi-role dialogue (Falk, 2026). There is a clear gap in

the development and adoption of assessment frameworks capable of tracking "compounding errors" and "lost in conversation" in real time (Measuring User Satisfaction, 2026).

2. **Second: Lack of Meta-Evaluation of Evaluators:** Despite the widespread adoption of modern evaluation frameworks based on linguistic models, previous research has rarely subjected the evaluator himself to rigorous systematic scrutiny (Guo et al., 2026). Using models as judges brings implicit biases that may mislead assessment, such as "position bias" and preference for lengthening, as well as a discursive (psychopathic) obedience to erroneous assumptions. The academic library lacks a standardized and comprehensive methodology for meta-evaluation that ensures the impartiality and consistency of the evaluator over time, a critical methodological gap that this research will highlight.

3. **Third: The Disconnect between Technical Assessment and Psychological Gap:** Current studies tend to address "technical response accuracy" and "user satisfaction" as separate research tracks. While some studies have looked at the effect of input interfaces (click-versus text interfaces) on user satisfaction based on the "expectation violation theory" (Tawseef & Kwon, 2025), this psychological dimension has not been combined with the technical degradation resulting from the "conversation tax" in complex interactions. Moreover, there is a gap in what is known as the "Empathy Paradox", where chatbots are still considered less empathetic than humans despite their superior language quality (The Illusion of Empathy, 2025).

How current research bridges this gap : The current research seeks to bridge the central gap between the rigorous technical and psychological dimensions of evaluating chatbots in modern web platforms. By building a standardized evaluation model that goes beyond superficial operational metrics, this study will correlate chatbot performance under multi-role conversations and deep semantic reliability (RAG Triad) to the psychological formation and actual satisfaction of users. In doing so, the research will provide a reliable and comprehensive scientific framework for platform developers that ensures a combination of high technical precision, impartiality, and outstanding human experience.

Chapter Three: Research Methodology

This chapter represents the procedural framework that translates the research objectives and questions posed in the first chapter into applicable scientific steps. This chapter explains the nature of the methodology, data collection tools, and analytical framework

used to evaluate chatbot performance and user satisfaction, ensuring the accuracy and reliability of the study outputs.

1. **Research Design**

Due to the complexity of the nature of the interactions between human users and large language model chatbots (LLMs), this study relied on **Qualitative Analytical Design** as the primary framework to guide the research path. This choice is due to the fact that sufficiency with purely quantitative and operational metrics (e.g., containment rates and response speed) does not provide a deep explanation of complex dialogical phenomena and variations in user psychological satisfaction (Stewart, 2025).

Psychologically, this approach provides space to analyze the mechanisms that shape user expectations based on input interfaces, and to interpret the gap between technical performance and predictions through the lens of Expectancy Violations Theory (Tawseef & Kwon, 2025). Technically, this design allows for deep contextual analysis of structural failures in multi-turn conversations, such as monitoring the causes of sycophancy, response inflation, and loss of diagnostic context over time (Falk, 2026; Guo et al., 2026). Furthermore, the research employs analytical methodology tools in the critical evaluation of advanced measurement frameworks and semantic scales, such as the RAG Triad and the use of linguistic models as judges (LLM-as-a-Judge) (Measuring User Satisfaction, 2026). Through this analytical approach, the study goes beyond simply monitoring numbers to assess the quality of benchmarks, their ability to detect "hallucinations" and absorb perceived empathy, ensuring that a comprehensive and reliable scientific framework that can be applied in modern web platforms is provided.

2. **Data Collection Tools & Sampling**

To achieve the objectives of the study and answer its research questions according to a qualitative-analytical design, this study relies on the integration of multiple data collection tools, which ensures an in-depth and comprehensive understanding of the performance of the chatbots and the level of user satisfaction. The main data collection tools are as follows:

1. **Conversation Logs & Transcript Analysis:** Text logs of actual conversations are the first and main tool for collecting technical data in this study. Transcripts of multi-turn conversations between users and chatbots will be extracted and analyzed to detect complex structural failures, such as premature answer attempts, answer bloated, and loss of context. The analysis of these records allows the evaluation of the accuracy of language models in preserving information, as studies indicate that the analysis of these complex interactions reveals an actual reduction in accuracy of up to 39%. To ensure the accuracy of the analysis, the Sliding window approach will be used when

reviewing the records, an analytical tool that ensures that the relevance of a response is assessed based on the cumulative context of the conversation and not just the roles immediately preceding it.

1. **User Satisfaction Questionnaires:** To collect data on psychological experience and the impact of input interfaces on the Theory of Violation of Expectations, the questionnaires will be used as a tool to collect user satisfaction data. These questionnaires will be designed to be presented to users immediately after the chatbot has completed their interaction, and will rely on the Likert scale (five-point or decimal) to rank satisfaction levels. Through this tool, quantitative scores and qualitative indicators will be extracted for the Customer Satisfaction Index (CSAT) and the Customer Effort Index (CES), as these metrics provide a first-hand understanding of the smoothness and effort exerted by the customer, and are an accurate indicator of the actual user experience.
2. **Automated LLM-as-a-Judge Tools:** Due to the large volume of texts and the complexity of semantic groundedness, advanced automated assessment tools will be employed to collect accuracy data as a supporting research tool. This tool is based on the LLM-as-a-Judge framework, which processes inputs (user question, retrieved context, and actual response) based on predefined criteria. This tool helps to extract accurate assessments of the RAG Triad, while providing textual explanations that explain why the answer is less accurate or correlated with reality, providing rich raw material for qualitative analysis.

1. Population and Sample

Due to the dual nature of the objectives of this research, which combine the measurement of psychological impressions (user satisfaction) and the analysis of software efficiency (response accuracy), the study population consists of two integrated parts: the human user community, and the language model community and technical conversation logs. In order to ensure an accurate representation of the reality of interaction in modern web platforms, the "Purposive Sampling" method was adopted to select the study vocabulary in both sections, as follows:

The human population of the study consists of all individuals who interact with AI-powered chatbots in digital platforms, and it is a huge and fast-growing community, as statistics indicate that the number of users of these systems reached more than 98 million users in 2022, and it is expected to rise to reach 110.9 million users by 2026. From this broad community, an intentional sample of users who have recently interacted with chat

agents in diverse fields (e.g., e-commerce, financial services, or technical support) will be drawn. This sample will be subjected to surveys aimed at measuring the Customer Satisfaction Index (CSAT) and the Customer Effort Index (CES). The intentional sample was chosen here to ensure that the input patterns they use (open text input versus click-based input) are different in order to effectively test the "expectation violation theory".

Second: Sample Language Models & Conversational Logs Sample For the technical aspect, the study population consists of all chatbots based on natural language processing in the web. To ensure the reliability and up-to-date of the analysis, the study sample consists of a set of Frontier Models that power the vast majority of current web platforms, including closed-source and open-source models such as: (GPT-4o), Claude 3.5 Sonnet, Gemini 1.5 Pro, and Llama 3.1 405B, as well as complex thinking models such as o1-preview.

The raw material analyzed (textual content) is a sample of "multi-turn conversations" logs. These records will be pulled from standardized databases approved for evaluation, such as the LLMEval 2 dataset of 2553 conversation samples paired with documented human preferences. The sample will also include lengthy conversations of up to 10 consecutive conversational roles. The purpose of selecting this complex sample of texts is to monitor structural deterioration in model accuracy, and to test critical points of failure such as "premature answer attempts", "answer bloat", and the "lost-in-middle effect".

Results and Discussion

1. **Results / Findings**

This section presents the quantitative and qualitative results obtained through the analysis of conversation logs and user satisfaction questionnaires, and these results are categorized into four main axes that reflect the objectives of the study:

First: User satisfaction and the impact of input patterns The quantitative results revealed a clear gap between the rates of institutional adoption of chatbots and the actual satisfaction of users. While 58% of companies rely on these systems, only 37% of users indicated that they were satisfied with the accuracy of the responses provided (Tawseef & Kwon, 2025). Furthermore, 63% of customers report that their recent interaction with chatbots has failed to resolve their underlying problem (Rajni, 2026).

Regarding the effect of input patterns according to the Expectation Violation Theory (EVT), the data showed that click-based input interfaces generate very high predictions in users, making any inaccurate response lead to severe "negative violation" and extreme frustration (Tawseef & Kwon, 2025),,. In contrast, users showed less clear expectations when interacting via "text-based", allowing for a "positive violation" that significantly boosted satisfaction levels when the bots provided accurate responses (Tawseef & Kwon, 2025),,. However, the "empathy paradox" emerged as a factor in satisfaction, with users rating chatbots (based on GPT models) as having high conversational quality and

technical superiority, but rated them as "less empathetic" tools compared to human customers (Measuring User Satisfaction, 2026).

Second: Response accuracy in multi-turn conversations Large language models recorded a sharp deterioration in performance when tested in a multi-turn environment compared to a single-turn environment. The average reduction in accuracy was 39% across all tested models (Falk, 2026). For example, the accuracy of the GPT-4.1 model decreased from 91.7% to 70.7%, and the accuracy of the Claude 3.7 Sonnet model decreased from 85.4% to 70.0% (Falk, 2026). More serious than the decline in accuracy is the deterioration in performance stability, with analyses showing that "Aptitude" has fallen by 16%, while "Reliability" has collapsed by 112% (Falk, 2026; Measuring User Satisfaction, 2026),.

Contextual analysis of multi-role conversations showed four main patterns of structural failure that led to this deterioration (Falk, 2026; Measuring User Satisfaction, 2026),,,:

1. **Premature attempts to answer:** Models were quick to generate answers during the first 20% of a conversation, resulting in an accuracy of only 30.9%, compared to 64.4% when waiting to gather the entire context.

2. **Lost-in-middle effect**: Language models ignored key information in the middle of conversations, and citation rates dropped to less than 20% (with a disregard rate of more than 80%).

3. **Answer inflation: The** length of robot responses increased across successive roles (roughly doubling from 700 to 1,400 characters in programming tasks) as a result of accumulating modifications to wrong assumptions rather than starting over.

4. **Accumulation of errors:** Failure to self-correct; once the model adopts an early erroneous conclusion, it continues to build on it silently in subsequent roles.

Fourth: Hallucinations & Groundedness Regarding the reliability of information, the overall average rate of "hallucinations" in the models was stable around 22% (Measuring User Satisfaction, 2026). The data showed that 55% of these hallucinations are caused by limitations in the data that feeds the robot (Measuring User Satisfaction, 2026). In terms of improvement, the results showed that reliance on "recall augmentation" (RAG) and structured indoctrination systems reduced hallucination rates by up to 22 percentage points, while non-RAG systems showed hallucination rates twice as high when dealing with precise or time-sensitive queries (Measuring User Satisfaction, 2026).

1. **Discussion**

This section provides an analytical interpretation of the findings, linking them to the theoretical framework and previous studies discussed in the second chapter, in order to directly answer the research questions raised in the first chapter.

First: Interpreting user satisfaction in light of the theory of expectation violation and the empathy paradox (answer to the first question) The results regarding the impact of input patterns are completely consistent with what he found (Tawseef & Kwon, 2025), as the data proved that option-based input interfaces (clicking) maximize users' preconceived expectations and lead to a severe "negative violation" when any error occurs, while open text interactions have a less clear expectation ceiling, allowing for a "positive violation" that enhances satisfaction when providing an answer Minute. Thus, we answer the first question that the input style is not just a UI tool, but a radical psychodeterminant of the customer's expectations. Furthermore, the results showed that the quality of the linguistic response does not necessarily imply emotional satisfaction, which is strongly in line with a study (Liu et al., 2025) on the "Empathy Paradox," which confirmed that chatbots, despite their technical and conversational superiority, are still evaluated by users as less empathetic than their human partners. This discrepancy clearly explains why actual customer satisfaction is low despite high rates of technology adoption in organizations.

Second: Structural Deterioration and Discursive Obedience in Multi-Role Conversations (Answers to Questions 2 and 3) Regarding the second question on the decline in accuracy in multi-role conversations (39%), our results are categorically consistent with the study (Falk, 2026) which confirmed that this decline is not just a transient deficiency, but rather the result of structural failures represented by "premature attempts to answer" before the context is complete, "answer inflation", and the effect of "lost-in-middle" Linguistic models ignore the critical context in the middle of a dialogue. As for the third question related to the impact of the "conversation tax" and the accumulation of errors, the results supported the conclusions (Guo et al., 2026). The analyses showed that language models suffer from severe weaknesses in "positive conviction" and tend to be docile (sycophancy) and go along with the user's erroneous suggestions rather than stick to the correct conclusion or safe refrain from answering. This finding answers the question of why reliability is collapsing, as robots fail to self-correct their course and accumulate early errors to reach misleading conclusions.

Third: The Effectiveness of Modern Evaluation Frameworks and Bridging the Gap (Answer to Question 4 and Question 1) to answer the fourth and main research question on how to evaluate and improve the performance of these systems, our discussion of the technical and psychological findings showed that traditional operational measures are no longer sufficient on their own. This view is consistent with the advanced evaluation literature that emphasizes the need to move beyond verbal conformity measures toward semantic reliability measures, such as the RAG Triad frameworkwhich breaks down the quality of the response to the accuracy of the recall, reliability, and relevance of the answer. Combining these deep technical metrics with psychological satisfaction metrics provides the unified and comprehensive framework that this research sought to build. Through this approach, developers can determine whether the failure is due to a "technical

hallucination" or a "negative violation of expectations," enabling them to optimize chatbots to be highly reliable and empathetic in modern web platforms.

Chapter V: Conclusion and Recommendations

Conclusion This study sought to deconstruct the complex relationship between the technical performance of large language model chatbots (LLMs) and the psychological impressions of user satisfaction within modern web platforms. The research concluded with a set of central conclusions:

First, user satisfaction is radically influenced by the pattern of the input interface: the "expectation violation theory" (EVT) has proven that click-based interfaces maximize user expectations, leading to extreme dissatisfaction when the bot fails to provide the desired response, while open text-based interfaces allow more room to positively exceed expectations and achieve higher levels of satisfaction.

Second, on the technical front, chatbots suffer from a sharp structural deterioration of up to a 39% reduction in response accuracy when moving to multi-turn conversations. This deterioration is due to complex technical phenomena, most notably the "lost-in-middle" effect, where the model ignores critical contexts, as well as premature haste in answering, and "answer inflation" through the accumulation of misinformation.

Third, the study revealed that the discursive reliability of language models is weak when subjected to dialectical pressure from the user, as systems tend to be submissive (sycophancy) and compromise their accurate diagnoses to conform to the user's erroneous suggestions rather than sticking to the correct conclusion. Moreover, these systems still suffer from the "Illusion of Empathy" (the paradox of empathy), as despite their superior language quality, users rate them as less empathetic and warmer than human elements, which is a barrier to full emotional satisfaction.

Finally, traditional operational metrics have been shown to fall short of assessing the efficiency of these systems, making the adoption of advanced semantic evaluation frameworks based on linguistic models as judges (LLM-as-a-Judge) – such as the RAG Triad – an operational necessity to ensure an accurate assessment of reliability, answer relevance, and the reduction of hallucinations.

Suggestions & Recommendations Based on the findings and conclusions, the study provides a set of practical recommendations for the industry, and suggestions for future research prospects:

First: Practical and Technical Recommendations

1. **Activate "Safe Abstention" policies: Web** platform developers recommend that chatbots be programmed and restricted to "safe abstention" mechanisms, so that systems acknowledge a lack of information rather than delusional or submissive to user errors, limiting the deterioration of reliability in long conversations.

2. **Adopt Hybrid Evaluation Methodologies:** Digital platforms should integrate automated evaluation tools (such as RAGAS and DeepEval) with Human-in-the-

loop monitoring to ensure that "silent failures" and subtle biases are detected that automated models cannot detect on their own.

3. **Designing interfaces that manage expectations:** UX designers are advised to carefully select input interfaces based on the bot's actual capabilities, so that open text interfaces are used to lower the ceiling of initial expectations and increase the likelihood of positive customer satisfaction when receiving an accurate answer.

Second: Proposals for Future Research Directions

1. **Meta-Evaluation of LLM Judges:** The study proposes to direct future research efforts towards developing standardized benchmarks for evaluating the evaluator himself, with the aim of measuring the biases inherent in the linguistic models used as judges, and examining their temporal stability to avoid what is known as "evaluation drift" over time.

2. **Explore MLLM-as-a-Judge: As conversational** agents evolve to accommodate images and sound, it has become necessary to conduct studies that measure the accuracy of response and user satisfaction in large multimedia language models (MLLMs), and investigate how these media affect perceived empathy.

3. **Specialization in High-Stakes Domains:** The study calls for the expansion of this research to apply to micro-sectors such as healthcare, financial services, and legal, as these environments require very stringent regulatory compliance standards that exceed the requirements of e-commerce platforms.

Third: Practical Steps to Measure and Improve Performance

To turn theoretical recommendations into measurable operational actions, it is recommended to adopt a periodic follow-up framework based on key performance indicators (KPIs) that combine efficiency, user experience, and operational impact. This approach ensures that it is not just about measuring response speed or usage volume, but also linking the evaluation of chatbots to actual resolution rates, satisfaction, and the ability to reduce the burden on support teams.

1. **Baseline Baseline Definition for 30–60 Days:** An organization begins by measuring the current state before any improvement, including automation rate, resolution rate from first contact, human conversion rate, user satisfaction score, and fallback rate.

Human Takeover ,Resolution Rate ,ah: Automation Rate'Musharat Al-Kafa .١
Average Handling Time ,Rate

2. User experience indicators: CSAT, CES, Return Visitor Rate, and indicators of negative behavior such as repetition or question paraphrasing.

3. Business Impact Indicators: Conversion Rate, Ticket Deflection Rate, and ROI.

2. **Establish clear operational thresholds:** Practical evidence suggests that good rates include automation between 70% and 85%, resolution rate over 65%, human conversion below 25%, CSAT above 80%, and Fallback below 10%. These limits can be used as initial benchmarks when evaluating a robot's performance in the actual environment.

3. **Confidence-Based Routing: Requests** beyond a high trust score are automatically processed, while low-trust requests are transferred to the human employee or to an additional validation layer, reducing the risk of hallucinations or unreliable answers.

4. **Review failure logs weekly: It** is recommended to extract the topics that caused the most fallback or human takeover and convert them directly into improvement elements in the knowledge base or in the design of the conversational flow. Practical practices show that analyzing the 10 most failed topics may achieve a significant reduction in failure rate and a rapid improvement in knowledge coverage.

5. **Integrate automated assessment with monthly human review: It** is not enough to rely on automated metrics alone, but periodic human review of conversation samples should be conducted to uncover issues of tone, loss of context, and conversational obedience, aspects that may not be fully reflected in abstract numbers.

6. **Create a unified dashboard:** Combines indicators of accuracy, satisfaction, task completion, containment, and human escalation, preventing optimizing one indicator at the expense of the rest of the experience. This integration helps detect situations where the robot appears to be digitally effective but actually impairs the user experience.

References

ACA International. (n.d.). *Banking chatbot study reveals security flaws as companies ramp up AI investment*.

Altmann, Y. (2026, March 23). *Chatbot KPIs 2026: The 10 most important metrics & benchmarks*. OMQ AI.

Ambos, A. (2026, April 12). *Top 10 AI chatbot customer service statistics 2026 shocking support automation breakthroughs*. Amra & Elma.

Armatis. (2025, September 26). *NPS, CES, CSAT… Which customer experience metrics should you choose?*

Barla, N. (2026, January 9). *Best RAG evaluation tools in 2026*. Adaline.

Bohn, N. (2026, March 5). *7 LLM evaluation & testing tools compared (2026)*. Rhesis AI Blog.

Buesing, E., Haug, M., Hurst, P., Lai, V., Mukhopadhyay, S., & Raabe, J. (2024, March 12). *Where is customer care in 2024?* McKinsey & Company.
Clayton, A. (2025, August 1). *Crowdtesting for AI systems in banks: Realistic tests for chatbots, voice systems and decision models*. Banking.Vision.
Cobbai. (2025, November 21). *Real-time customer sentiment analysis with AI: Tools and how-to guide*.
Consumer Financial Protection Bureau. (2023, June 6). *Chatbots in consumer finance*.
Delicious-One-5129. (2026). *Top LLM observability tools comparison I tried for agents in production* [Online forum post]. Reddit.
Dolan, E. W. (2026, May 9). *ChatGPT's free version is 26 times more likely to respond inappropriately to psychotic delusions*. PsyPost.
Falk, F. (2026, April 29). *Your AI agent loses 39% accuracy in real conversations. ICLR 2026's outstanding paper explains why*. Beam AI.
FeedbackRobot. (2025, October 14). *Using AI sentiment analysis to understand customer emotions in feedback for 2026 [AI Powered]*.
Genuity Systems Ltd. (2025, October 28). *50 key statistics on AI chatbots in customer service*.
Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., ... & Guo, J. (2024). A survey on LLM-as-a-judge. *arXiv*.
Guo, K. H., Yan, C., Baidya, A., Brown, K., Gao, X., Xiong, J., Yin, Z., & Malin, B. A. (2026). Stop listening to me! How multi-turn conversations can degrade LLM diagnostic reasoning. *arXiv*.
Intent mismatch causes LLMs to get lost in multi-turn conversation. (n.d.). ResearchGate.
Ip, J. (2026, May 12). *Top 5 LangSmith alternatives and competitors, compared (2026)*. Confident AI.
Kim, S. (2024). *Prometheus-eval: Evaluate your LLM's response with Prometheus and GPT4* [Computer software]. GitHub.
LangChain. (n.d.). *8 LLM observability tools to monitor & evaluate AI agents*.
Lewis, A. (2025, October 2). Detecting LLM hallucinations. *DataConnect Conference 2025*. Women in Analytics & Women in Data.
LiveAgent. (2026, January 20). *The top 16 customer service metrics to measure in 2025*.
Liu, T., Giorgi, S., Aich, A., Lahnala, A., Curtis, B., Ungar, L., & Sedoc, J. (2025). The illusion of empathy: How AI chatbots shape conversation perception. *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence (AAAI-25)*.
Measuring user satisfaction and response accuracy of NLP-based chatbots in modern web platforms. (2026).
P., Bram. (2026, January 30). *Top 8 LLM observability tools: Complete guide for 2025*. LangWatch.
Rajni. (2026, April 7). *30+ AI customer service statistics *. YourGPT Blog.

Rastegar-Panah, M. (2026, May 6). *Customer satisfaction scores (CSAT): what it is & how to measure*. Zendesk.
Sarha, S. (2024, December 12). *2024 recap: Measuring regulatory complexity, AI-generated compliance and more*. Apiax.
Shah, K. (2026, January 21). *Top 5 tools to evaluate RAG performance in 2026*. Maxim AI.
Sirdeshmukh, V., Deshpande, K., Mols, J., Jin, L., Cardona, E.-Y., Lee, D., ... & Xing, C. (n.d.). MultiChallenge: A realistic multi-turn conversation evaluation benchmark challenging to frontier LLMs. *ACL Anthology*.
Sivakolundhu, R., & Yagamurthy, D. N. (2024). Adaptive chatbots: Real-time sentiment analysis for customer support. *International Journal of Computing and Engineering (IJCE)*, 6(1), 55-64.
SQ Magazine. (n.d.). *LLM hallucination statistics 2026: AI gets facts wrong up to 82% of the time*.
Stewart, A. (2025, December 26). *7 conversational AI evaluation metrics that actually matter in 2025*. Dialzara.
Tawseef, T., & Kwon, W.-S. (2025). Clicks or text? How customer input mode and chatbots' response accuracy shape customer experience for fashion brands. *2025 Proceedings St. Louis, Missouri*. International Textile and Apparel Association.
Technological folie à deux: Feedback loops between AI chatbots and mental health. (n.d.). PubMed Central (PMC).
Vongthongsri, K. (2025, October 10). *G-Eval simply explained: LLM-as-a-judge for LLM evaluation*. Confident AI.
Winks, E. (2026, April 10). *RAGAS vs. TruLens vs. DeepEval: LLM evaluation frameworks compared | A 2026 guide*. Atlan.