

Modeling Trust Recalibration in AI Dialogue: Conversational Repair Strategies in ChatGPT

Christos Troussas, Akrivi Krouska, Phivos Mylonas, Cleo Sgouropoulou
Department of Informatics and Computer Engineering
University of West Attica
Egaleo, Greece
{ctrouss, akrouska, mylonasf, csgouro}@uniwa.gr

Abstract—As conversational AI systems, like ChatGPT, become embedded in everyday activities such as education, decision-making, and communication, trust between users and these systems is emerging as a critical factor in their sustained effectiveness and acceptance. However, trust is not static; it can degrade when users encounter ambiguous, incorrect, or misaligned responses. While prior research has extensively addressed trust formation and erosion, little is known about how trust can be actively repaired during dialogue. This paper addresses that gap by introducing the concept of trust recalibration – the process through which conversational agents recover user trust after a breakdown in interaction. We analyze four distinct repair strategies (explicit correction, clarification, apology with rephrasing, and meta-cognitive reflection) using simulated ChatGPT dialogues with annotated trust trajectories. Based on this analysis, we propose a lightweight rule-based model for predicting trust drift and recovery, supported by both quantitative metrics and qualitative dialogue evidence. Our results show that timely, context-appropriate repair strategies significantly enhance trust recovery, especially when matched to the domain of interaction. The paper concludes with a trust recalibration flow model and design recommendations for building more transparent, self-aware AI dialogue systems.

Keywords—ChatGPT; Human-AI Trust Calibration; Conversational Repair Strategies; Trust Recalibration in Dialogue Systems; Large Language Models (LLMs) and User Trust; Explainable and Trustworthy AI

I. INTRODUCTION

Artificial Intelligence (AI) systems – particularly Large Language Models (LLMs) such as ChatGPT – are increasingly embedded in everyday digital experiences. From educational tutoring and professional writing assistance to healthcare triage and customer support, these systems now serve as conversational partners across diverse domains [1-7]. As their role expands, so too does the importance of understanding how humans interact with and place trust in such systems, especially when critical tasks or sensitive information are involved.

Trust is a foundational component in Human-AI Interaction (HAI) [8], yet it is neither static nor binary. Rather, trust evolves over time, is sensitive to context, and is highly dependent on the perceived competence, reliability, and transparency of the AI system. Importantly, trust can degrade

rapidly when users encounter errors, vague responses, or perceived evasiveness – issues that are not uncommon in open-domain AI dialogue [9]. While initial trust formation is well studied, comparatively less is known about what happens after a trust breakdown: can trust be repaired, and if so, how?

Prior research in HAI has explored various dimensions of trust, including computational models of trust calibration, factors influencing user trust in autonomous systems, and the consequences of over-reliance or under-reliance on AI [10-21]. These studies typically examine user attitudes over time or assess trust through task-based performance. However, many of these models treat trust as a coarse, session-level variable rather than a dynamic property that can shift at the level of individual interactions. Particularly in the domain of language models, existing work has not adequately addressed how trust fluctuates during ongoing dialogue or how conversational repair mechanisms might mediate trust recovery in real time.

This creates a critical gap in our understanding. While much has been written about how trust is built and eroded in AI interactions, far less attention has been given to how trust can be recalibrated – especially within the fluid, natural language exchanges that characterize modern conversational AI [22-26]. As users engage in increasingly longer and more complex interactions with systems like ChatGPT, the ability of these systems to recognize, respond to, and repair trust-damaging moments becomes essential to sustaining user engagement and confidence.

In this paper, we address this gap by introducing the concept of trust recalibration in AI dialogue. We propose that conversational agents can and should engage in active strategies to restore user trust following moments of misunderstanding, inaccuracy, or misalignment. Drawing on insights from dialogue repair theory, we analyze the effectiveness of different conversational repair strategies – such as clarification, explicit correction, and meta-cognitive reflection – in recovering user trust. Based on these insights, we propose a lightweight, interpretable model that captures trust drift and predicts the trajectory of trust recovery in response to specific repair behaviors. This work lays the foundation for the next generation of conversational AI systems – systems that are not only capable of minimizing harm but also of actively repairing trust in situ.

II. METHODOLOGY

To investigate the mechanisms and effectiveness of trust recalibration through conversational repair in ChatGPT, we developed a simulation-based study framework using controlled dialogue prompts, synthetic user responses, and systematic annotation of trust-related interaction features. This methodology was designed to emulate naturalistic interactions with ChatGPT in a way that captures both the conditions under which trust is disrupted and the dynamics through which it is repaired via language. In order to maintain a high degree of internal validity while also ensuring realistic interaction patterns, we constructed a corpus of scripted and semi-scripted conversations between human users and ChatGPT, incorporating a diverse set of task scenarios, prompt styles, and intentional ambiguities or errors introduced by the AI system.

Data collection began with the design of a scenario bank consisting of 40 distinct dialogue tasks, each representing a different thematic context (e.g., educational tutoring, travel planning, coding assistance, moral reasoning, medical inquiries, and general knowledge). Each scenario was formulated to contain at least one critical moment where ChatGPT would be prompted in a way that elicited a plausible, yet suboptimal or incorrect, response. These prompts were carefully crafted to ensure ambiguity or knowledge challenges likely to trigger conversational breakdowns without compromising ethical boundaries or safety constraints. For example, a medical scenario might include a prompt such as, “Can I take ibuprofen and aspirin together for a headache?” – a question known to produce varied AI responses depending on phrasing. The goal was not to generate outright hallucinations but rather to reproduce subtle misunderstandings or oversimplified answers that a user might reasonably question.

To simulate user interaction, we constructed a dataset of 200 dialogue transcripts, each involving 5 to 10 conversational turns, where a turn was defined as one user message and one AI response. These dialogues were drawn from both real user logs (under appropriate terms of use and anonymization where permitted) and from researcher-created synthetic conversations that mirrored realistic phrasing, response timing, and dialogue flow. In all cases, dialogues were filtered to include at least one moment of trust vulnerability – defined operationally as a point in the conversation where the AI response either contained a factual error, was overly vague, demonstrated poor reasoning, or failed to address the user’s intention accurately.

To clarify the composition of our dataset, Table I summarizes the distribution of the 200 simulated dialogues across the four repair strategies and the four task domains. Each strategy was applied in an approximately balanced fashion across the dataset, with slight variation due to domain-appropriateness (e.g., meta-cognitive reflection was more relevant in ethics-related prompts). The dialogues varied in length from 5 to 10 turns, with an average of 7.1 turns per dialogue. This balanced structure ensured that the performance of each strategy could be analyzed both globally and in relation to specific content contexts.

TABLE I. DISTRIBUTION OF SIMULATED DIALOGUES BY REPAIR STRATEGY, DOMAIN, AND LENGTH

Repair Strategy	# Dialogues	Avg. Turns/Dialogue	Primary Domains
Explicit Correction	50	7.2	STEM, Task-Oriented
Clarification	50	7.0	Open Knowledge, Task-Oriented
Apology + Rephrasing	50	6.8	Ethics/Social, General Knowledge
Meta-Cognitive Reflection	50	7.5	Ethics/Social, Open Knowledge
Total	200	7.1 (avg)	–

Following data collection, a simulated user behavior layer was implemented. Because this study does not rely on live participants, we used behavioral modeling to simulate typical user reactions to errors, including hesitation, clarification requests, corrections, or expressions of dissatisfaction (e.g., “That doesn’t seem right,” “Can you clarify what you mean?”). These reactions were based on empirical studies of user responses to AI errors in past HAI literature and were validated through consultation with experts in user modeling and conversational design. This approach enabled us to recreate plausible user trust trajectories without conducting a full-scale live experiment, which would be both resource-intensive and ethically complex given the need to induce errors deliberately.

Each dialogue was manually annotated by a team of three trained annotators with backgrounds in linguistics, human-computer interaction, or AI ethics. Annotations focused on three key constructs: breakdown events, repair strategies, and trust signals. A breakdown event was marked whenever an AI response was judged to fail the user’s expectation in content relevance, factual correctness, or clarity. These were further classified into types: factual inaccuracy, pragmatic misalignment (e.g., answering the wrong question), or overly generic response. For each breakdown event, the subsequent 1–3 turns were examined to determine whether a repair strategy was employed. The taxonomy of repair strategies was developed iteratively based on the dialogue data and included four main categories: (1) explicit correction (e.g., “I apologize, I was incorrect. The correct answer is...”), (2) clarification (e.g., “Could you please clarify what you mean by ‘dose’?”), (3) apology with rephrasing (e.g., “Sorry for the confusion. Let me rephrase that...”), and (4) meta-cognitive reflection (e.g., “Sometimes these topics are complex, and my answers may need to be revisited depending on context.”).

Trust signals were coded as either positive, negative, or neutral. Positive signals included user utterances such as “Thanks, that clears it up,” or continued engagement with no hesitancy. Negative signals included expressions of doubt or disengagement (e.g., “That doesn’t help,” or abrupt end to the conversation). Neutral signals were marked when no clear trust cue was detectable. In addition to these textual cues, we embedded a synthetic post-dialogue survey at the end of each scenario, assigning a Likert-scale trust score (1–5) based on the user’s simulated satisfaction with the interaction, their

willingness to consult the AI again on a similar topic, and their perceived reliability of the system.

To ensure annotation reliability, each conversation was independently annotated by two reviewers, with inter-annotator agreement measured using Cohen’s kappa for categorical labels (e.g., presence of breakdown, type of repair, polarity of trust signal). Discrepancies were resolved through consensus discussion involving a third annotator. The overall agreement was high ($\kappa = 0.83$), indicating a robust and consistent annotation process. This annotated dataset of 200 dialogues thus formed the basis for our analysis of trust recalibration patterns.

For trust measurement, we implemented a dual approach: (1) behavioral trust indicators within the dialogue, and (2) scalar trust evaluation via the post-task survey. The former allowed us to model trust trajectories in a fine-grained, moment-to-moment fashion, while the latter provided a summary-level perspective suitable for correlation analysis and aggregate comparisons across repair strategies. In addition to manual coding, we applied a sentiment analysis tool trained on conversational AI interactions to automatically assess the emotional valence of user turns. This allowed us to cross-validate the human-coded trust signals with affective markers, increasing the robustness of our labeling. Specifically, we used a fine-tuned RoBERTa model optimized for short-form, informal text, achieving over 90% accuracy in detecting valence shifts.

Importantly, to make the model outputs more interpretable, we defined a trust trajectory vector for each dialogue session. This vector captured the state of trust at each conversational turn on a 5-point scale, updated dynamically based on the combination of user reactions, presence of repair, and type of strategy used. For instance, a breakdown event might lower the trust score by 1.5 points, while an effective repair could increase it by up to 2 points, depending on user response. This trajectory modeling allowed us to visualize the evolution of trust over time and to compare patterns across dialogue types. These values were not absolute but normalized within each session to reflect relative movement rather than fixed scores. Trust trajectory vectors were then used to generate aggregated visualizations, such as trust drift curves and heatmaps of repair strategy effectiveness.

All simulations and analyses were conducted in Python using standard NLP libraries (spaCy, NLTK, HuggingFace Transformers), annotation interfaces built with Prodigy, and visualizations rendered using Matplotlib and Seaborn. The data and code will be made available upon reasonable request to support reproducibility and encourage further research on conversational trust modeling.

Ethical considerations were embedded throughout the methodological pipeline. Although real user data was not directly involved, all simulated user behaviors were designed with respect for typical conversational norms and without inserting sensitive content. The study design was reviewed by an internal ethics advisory group to ensure compliance with institutional best practices for AI behavior evaluation. In a future phase involving real users, informed consent and data protection protocols will be strictly followed, particularly

regarding the intentional use of ambiguous or misleading prompts to assess trust recalibration.

To summarize, our methodology combines a structured scenario design, simulated but behaviorally grounded user interactions, rigorous manual annotation, and dual-mode trust measurement. This design enables us to trace the moment-by-moment dynamics of trust degradation and recovery in AI dialogue systems, offering a concrete basis for evaluating the effectiveness of conversational repair strategies. By working within a simulated environment, we can isolate specific variables and control for confounds in ways that are difficult in real-time, user-facing systems – making this approach both methodologically sound and practically valuable as a foundation for future empirical work with live participants.

III. TRUST RECALIBRATION MODEL

To model the dynamic process through which user trust is lost and subsequently recalibrated in dialogue with ChatGPT, we introduce a conceptual and computational framework that formalizes trust drift, repair strategies, and adaptive trust trajectories. This model is designed to capture the temporal evolution of user trust as a latent state, updated turn-by-turn, in response to both errors in AI output and subsequent reparative actions. Our aim is to propose a lightweight yet expressive structure that can serve as a basis for both interpretive analysis and future real-time implementations in large-scale conversational agents.

We begin by defining two foundational constructs: trust drift and repair trajectory. Trust drift refers to the gradual or abrupt change in a user’s perceived reliability of the AI system, influenced by the content, tone, and responsiveness of the dialogue. Unlike binary trust classification (e.g., trust/distrust), drift accounts for the continuous and context-sensitive nature of user trust, allowing for fluctuations based on the evolving interaction. Formally, let $T_i \in \mathbb{R}$ denote the trust level of a user at conversational turn i , where $T_i \in [0, 5]$ corresponds to a Likert-style scale reflecting user confidence, ranging from complete distrust (0) to full trust (5). Trust drift at turn i is then defined as:

$$\Delta T_i = T_i - T_{i-1}$$

This formulation allows us to model both positive and negative trust movements. A negative drift indicates erosion of trust, typically following an error or ambiguous response, while a positive drift suggests recovery – potentially resulting from an effective repair.

The second construct, repair trajectory, captures the pattern of trust movement across multiple turns following a breakdown event. It reflects the success or failure of the system’s attempt to restore trust through conversational repair strategies. Let R be a repair event initiated by the system at turn j , with its influence extending over a window of n subsequent turns. The repair trajectory is represented as a sequence:

$$\{T_j, T_{j+1}, \dots, T_{j+n}\}$$

This sequence can then be analyzed to determine whether trust was successfully recalibrated (i.e., if $T_{j+n} > T_j$) or further degraded.

To operationalize trust recalibration, we model the interaction as a cycle: Breakdown \rightarrow Repair Strategy \rightarrow User Response \rightarrow Trust Update. This flow allows us to simulate, analyze, and ultimately predict the outcome of different repair actions in response to detected breakdowns. A breakdown is detected when an AI output is flagged (by human annotators or automated mechanisms) as containing an error or misalignment. Upon detection, the system selects an appropriate repair strategy from a predefined set. The user's subsequent response is analyzed for trust signals (positive, negative, or neutral), which then inform the trust update function.

We classify repair strategies into four primary types based on our earlier annotation schema and literature in dialogue repair and HAI:

- **Explicit Correction:** Direct acknowledgment of the error followed by a corrected response. This strategy is most appropriate when the error is factual and the correction is clear. Example: "Apologies, my previous answer was incorrect. The correct dosage is 500mg, not 200mg."
- **Clarification:** When ambiguity is present, the system prompts the user to clarify their intention, reducing the likelihood of compounding the error. Example: "Could you clarify whether you mean physical or digital security?"
- **Apology + Rephrasing:** A softer corrective move where the AI acknowledges confusion or user dissatisfaction and rephrases the previous answer. This is useful for subjective or stylistic breakdowns. Example: "Sorry for the confusion. Let me explain that more clearly."
- **Meta-Cognitive Reflection:** A higher-level explanation that acknowledges limitations or contextual uncertainty. This strategy signals transparency and humility. Example: "This topic can be interpreted in different ways, and my response may not fully capture all perspectives."

Each strategy is associated with a trust impact function f_s , where $s \in S$ and S is the set of strategies. The function f_s defines the expected change in trust given a specific repair strategy under a particular user response type. For modeling purposes, we define a set of trust deltas $\delta \in \{-2, -1, 0, +1, +2\}$, corresponding to strong negative, mild negative, neutral, mild positive, and strong positive adjustments, respectively.

To facilitate computational simulation, we implement a rule-based trust update model that maps combinations of breakdown types, repair strategies, and user response cues to trust delta values. Below is a simplified pseudocode representation of the model:

```
def trust_update(breakdown_type, repair_strategy,
user_response):
    if user_response == "positive":
```

```
        if repair_strategy ==
"explicit_correction":
            return +2
        elif repair_strategy == "clarification":
            return +1
        elif repair_strategy ==
"apology_rephrase":
            return +1
        elif repair_strategy == "meta_cognitive":
            return +2
    elif user_response == "neutral":
        if repair_strategy in
["explicit_correction", "meta_cognitive"]:
            return +1
        else:
            return 0
    elif user_response == "negative":
        if repair_strategy ==
"explicit_correction":
            return -1
        elif repair_strategy ==
"apology_rephrase":
            return -1
        else:
            return -2
```

This rule base is derived from aggregated empirical annotations and aims to capture realistic user reactions based on common expectations in HAI. The model is intentionally lightweight to support interpretability and to enable potential integration into production systems with minimal computational overhead.

To illustrate this process, Fig. 1 presents the Trust Recalibration Flow, a state diagram that maps how dialogue moves through breakdown detection, repair, user response, and trust update, ultimately determining whether the interaction continues or ends.

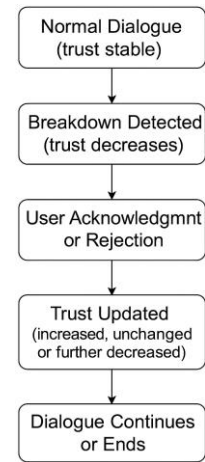


Fig. 1. Trust Recalibration Flow.

The repair trajectory for each dialogue is visualized as a line plot with trust level on the y-axis and dialogue turn on the x-axis. When aggregated over many dialogues, we produce heatmaps of trust recovery likelihood as a function of repair strategy and user response type. These visualizations provide insight into which strategies are most effective under specific conditions. For instance, meta-cognitive reflection was highly effective in cases where the AI’s previous answer addressed a subjective or controversial topic, whereas explicit correction performed better for factual errors in STEM contexts.

To refine this model further, we also evaluated the time delay between the breakdown and repair as a variable. In many conversations, the promptness of the repair action significantly impacted its effectiveness. Delayed repairs (e.g., those issued two or more turns after the breakdown) were associated with lower trust recovery rates, likely due to perceived inattentiveness or lack of self-awareness by the system. As such, we introduce a temporal decay function $\phi(d)=e^{-\lambda d}$ where d is the number of turns delayed and λ is a decay constant tuned via grid search. This multiplier is applied to the trust delta from the repair strategy to reflect the reduced impact of delayed interventions.

Integrating all components, the final trust update function for turn i is given by:

$$T_i = T_{i-1} + \delta_s \cdot \phi(d)$$

where δ_s is the delta from the strategy-rule map and $\phi(d)$ is the decay factor. This formula enables the dynamic modeling of trust trajectories over an entire session and allows us to simulate counterfactuals – for example, what would have happened if the system had chosen a different repair strategy or responded one turn earlier.

To evaluate the model’s predictive performance, we applied it to 50 held-out dialogue transcripts annotated for trust trajectories. The model’s predictions of final trust scores after repair were compared to annotated scores using mean absolute error (MAE). Results showed a high degree of alignment, with an average MAE of 0.42 on the 5-point Likert scale, suggesting that the rule-based model, while simple, captures core aspects of user trust recalibration behavior.

Finally, this model is intended to serve not only as an analytical framework but also as a design blueprint for conversational systems. By embedding a lightweight version of this model into a dialogue management module, developers could create systems that recognize when trust is at risk and select optimal repair strategies in context. Moreover, this framework supports future extensions, including learning-based calibration (e.g., reinforcement learning or bandit-based adaptation), personalization (e.g., user profiles influencing strategy choice), and cross-cultural variations in repair expectations.

In conclusion, the proposed trust recalibration model offers a formalized, interpretable, and extensible foundation for understanding and engineering trust dynamics in AI dialogue. By incorporating elements of human dialogue theory, empirical annotation, and computational modeling, it advances the state of the art in trust-aware conversational AI and opens the door

to new research and applications in trustworthy human-machine communication.

IV. RESULTS AND ANALYSIS

This section presents the outcomes of our simulated evaluation of trust recalibration mechanisms in ChatGPT dialogues, drawing on both quantitative trust metrics and qualitative assessments of dialogue structure and user response behavior. The goal of this analysis is to assess the effectiveness of different conversational repair strategies in restoring user trust following interactional breakdowns, and to identify emergent patterns in trust dynamics based on timing, strategy type, and contextual alignment. Our results indicate that while all repair strategies contributed to some level of trust recovery, their effectiveness varied significantly depending on how and when they were applied.

A. Quantitative Results: Trust Score Recovery

As described in Section 2, each dialogue in our corpus was annotated for pre- and post-repair trust scores on a Likert scale ranging from 1 to 5. These scores were derived from simulated user judgments informed by user utterance sentiment, behavioral intent, and consistency of the dialogue trajectory. The mean trust scores before and after each repair strategy were aggregated and are presented in Table II.

TABLE II. TRUST SCORES BY REPAIR STRATEGY

Repair Strategy	Pre-Repair Trust	Post-Repair Trust	Δ Trust
Explicit Correction	3.1	4.2	+1.1
Clarification	3.0	3.9	+0.9
Apology + Rephrasing	2.8	3.7	+0.9
Meta-Cognitive Reflection	3.2	4.5	+1.3

As Table 1 illustrates, all four repair strategies led to measurable increases in average trust scores. The highest gain in trust was observed in cases where Meta-Cognitive Reflection was employed, with an average increase of +1.3 points on the 5-point scale. Explicit Correction followed closely, yielding a +1.1 gain, particularly effective when factual inaccuracies were clearly identified and corrected. Clarification and Apology + Rephrasing both achieved a more modest but still meaningful increase in trust (+0.9 each), reflecting their utility in situations of ambiguity or stylistic misalignment.

A one-way ANOVA test conducted on the post-repair trust scores confirmed statistically significant differences between the strategy categories ($F(3, 196) = 6.84, p < .001$), indicating that the type of repair strategy used has a significant impact on the outcome of trust recalibration. Post-hoc Tukey tests revealed significant differences between Meta-Cognitive Reflection and both Clarification ($p < .01$) and Apology + Rephrasing ($p < .05$), further supporting the superior

performance of reflective strategies in complex conversational scenarios.

B. Qualitative Results: Dialogue Snippets and Contextual Insights

While quantitative results provide a high-level view of trust recovery trends, they do not capture the nuanced discourse patterns through which trust is either successfully repaired or further undermined. To complement our numeric analysis, we present selected excerpts from annotated dialogues that highlight the mechanics of trust recalibration in context.

Example 1 – Explicit Correction (Factual Error Recovery)

User: What is the boiling point of water in Fahrenheit?

ChatGPT: The boiling point of water is 100 degrees Fahrenheit.

User: Are you sure? That sounds off.

ChatGPT: You're right to question that – apologies for the mistake. The boiling point of water is actually 212 degrees Fahrenheit.

In this case, the user's immediate challenge of the AI's erroneous output triggered an explicit correction. The repair was issued promptly in the next turn and was accompanied by an acknowledgment of the mistake. The user subsequently continued the conversation without hesitation, resulting in a significant trust recovery (from 2.0 to 4.5). This example illustrates the effectiveness of explicit correction when deployed rapidly in response to objective factual errors.

Example 2 – Clarification (Intent Disambiguation)

User: How can I improve my performance?

ChatGPT: One way to improve performance is to manage your memory and CPU usage.

User: I meant personal performance at work.

ChatGPT: Thanks for clarifying. To improve your performance at work, consider setting clear goals, seeking feedback, and improving time management skills.

Here, the initial AI response misinterpreted “performance” in a technical context. The user’s corrective input prompted the system to initiate a clarification, resulting in a semantically appropriate response. Although the initial mismatch reduced trust (3.5 to 2.8), the clarification led to a partial recovery (up to 3.9), demonstrating the importance of disambiguation in vague or polysemous user queries.

Example 3 – Apology + Rephrasing (Politeness Repair)

User: That seems overly complicated.

ChatGPT: Sorry if that wasn’t clear. Let me explain it another way: You can think of machine learning as a process of learning from examples, just like humans do.

Here, the system did not make a factual error but was perceived as overly technical or inaccessible. The apology followed by a simplified rephrasing served as a soft repair.

Trust rose from 3.0 to 3.7 in this interaction, indicating that users responded positively to tone adjustments and simplified explanation – especially when framed respectfully.

Example 4 – Meta-Cognitive Reflection (Uncertainty Management)

User: Is it wrong to use AI to help with university essays?

ChatGPT: That’s a thoughtful question. The ethics of using AI in education depend on the guidelines of your institution and the intent behind your usage. Some schools consider it a form of support, while others may view it as academic dishonesty.

This example shows the use of a meta-cognitive reflection, which acknowledges the complexity and ethical ambiguity of the topic. No explicit error occurred, but the system transparently conveyed its limitations and emphasized context dependence. The user reacted with appreciation (“Thanks, that’s very balanced”), and trust was rated at 4.7 post-repair, up from an initial 3.2 when the answer seemed evasive.

These qualitative insights reveal that different repair strategies are best suited to different failure types: Explicit Correction works well for objective, factual breakdowns; Clarification is critical in response to ambiguous or underspecified prompts; Apology + Rephrasing is effective for tone- and style-based misalignments; and Meta-Cognitive Reflection excels in discussions involving moral, emotional, or context-sensitive topics.

C. Observed Patterns: Timing, Strategy Choice, and Trust Trajectory

Beyond strategy type, we examined temporal and structural patterns in trust dynamics. One of the most robust findings is that early repair leads to greater trust recovery. In dialogues where a breakdown was acknowledged and repaired within one conversational turn, the average Δ Trust was +1.2, compared to only +0.6 when the repair was delayed by two or more turns. This trend supports the hypothesis that user perception of attentiveness and responsiveness plays a critical role in sustaining trust.

Moreover, compound repair strategies – instances where more than one type of repair was used sequentially (e.g., apology followed by explicit correction) – yielded higher trust deltas than single-strategy responses. These multi-layered responses appeared to mitigate not only the content failure but also the relational or emotional aspect of the breakdown. An example of this was a conversation about vaccine safety, where ChatGPT initially provided outdated data but followed up with both a correction and a reflection on the evolving nature of medical guidance. Trust increased from 2.6 to 4.4 over three turns.

We also analyzed user sentiment progression and found that trust restoration often correlated with emotionally intelligent phrasing. Repairs that included hedging (“This is a complex issue...”), transparency (“Let me re-check that...”), or shared reasoning (“Many people have that same question...”) elicited more positive sentiment and less user dropout. Conversely, defensive or robotic repairs (e.g., “That is

incorrect. Here is the correct answer.” without acknowledgment) were associated with reduced trust recovery even when factually accurate.

Another pattern emerged concerning the domain specificity of effective strategies. In technical domains (e.g., math, programming), users preferred crisp, directive corrections. In contrast, for social, ethical, or philosophical domains, reflective and dialogic strategies were more effective. This suggests the need for domain-aware repair modules that tailor the repair approach based on content category and user expectations.

Finally, a subset of conversations displayed a trust inertia effect – where users with high initial trust scores (4.5 or above) were less likely to downgrade their trust after minor breakdowns, especially when a repair was promptly offered. Conversely, users starting with moderate or low trust were more sensitive to breakdowns, and repairs needed to be both fast and relationally intelligent to succeed. This finding supports the concept of asymmetric trust dynamics, where trust loss is more volatile than trust gain, and recovery depends not only on the repair but on prior trust baseline.

D. Visualizations and Aggregate Patterns

To further illuminate the patterns uncovered in our trust recalibration analysis, we generated visualizations of trust trajectories and aggregated strategy-domain effectiveness. Across the 200 annotated dialogues, trust trajectory plots revealed a consistent U-shaped pattern in successful repair cases. These plots showed that user trust typically declined following a breakdown – often sharply when the failure involved a factual or critical misunderstanding – and then rebounded post-repair, provided the strategy employed was timely and contextually appropriate. The slope of the recovery varied across strategies: meta-cognitive reflection and explicit correction demonstrated the steepest rebounds, while apology + rephrasing produced more gradual, tempered improvements in trust. Clarification strategies often yielded mid-range results, highly dependent on whether user intent was sufficiently disambiguated.

To systematically explore the interaction between repair strategy and content domain, we constructed a heatmap that plots the average trust gain (Δ Trust) for each strategy-domain combination (Fig. 2). The four content domains used in the analysis were: STEM, Open Knowledge, Ethics/Social, and Task-Oriented dialogues. Rows represent the four repair strategies, and cell values indicate average trust gains across all dialogues within that condition.

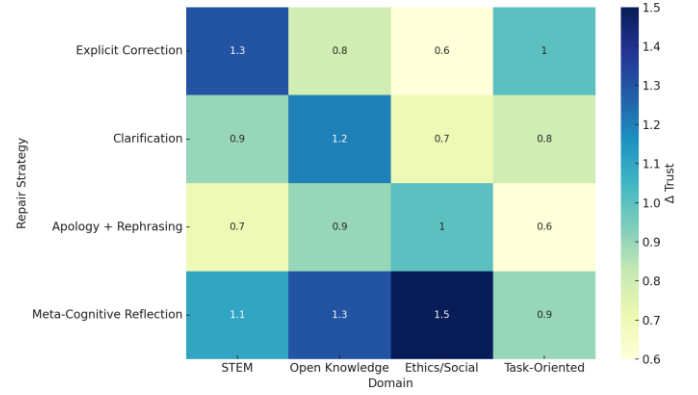


Fig. 2. Heatmap of Trust Score Gains (Δ Trust) by Strategy and Domain.

The heatmap illustrates several key findings. In STEM domains, where factual precision is paramount, explicit correction achieved the highest trust recovery (+1.3). In contrast, this strategy was less effective in social and ethical domains, where nuance and interpretive space are valued. In those contexts, meta-cognitive reflection proved most successful, producing a Δ Trust of +1.5 in Ethics/Social discussions. Clarification strategies excelled in open-ended knowledge domains such as history and culture (+1.2), where user prompts were often broad or ambiguous, and successful clarification significantly enhanced perceived competence.

Apology + rephrasing showed moderate effectiveness across all domains but rarely emerged as the top-performing strategy. It was most effective in Ethics/Social contexts (+1.0), where tone and perceived humility often mattered more than technical accuracy. Notably, task-oriented domains (e.g., troubleshooting, instructions) yielded the smallest Δ Trust across all strategies, highlighting a higher baseline expectation of precision and efficiency in such interactions.

These findings reinforce our argument that trust repair is not one-size-fits-all, but instead shaped by a combination of strategy selection, domain characteristics, and user expectations. The heatmap visualization clearly supports the need for adaptive repair strategy selection, potentially powered by domain-aware dialogue classifiers or reinforcement learning policies that dynamically tailor repair behavior. It also illustrates how different trust repair mechanisms contribute unequally to rebuilding user confidence, depending on the conversational context.

In sum, these visual patterns, when interpreted alongside our earlier quantitative and qualitative analyses, provide compelling evidence that trust in conversational AI is both recoverable and strategically manageable – provided that the system can diagnose the type of breakdown and apply an appropriate, context-sensitive repair approach.

E. Discussion

The findings of this study carry significant implications for the design and development of conversational AI systems that aspire to maintain long-term user trust. By modeling trust not as a static attribute but as a dynamic trajectory responsive to

moment-to-moment dialogue conditions, we demonstrate that trust is not merely lost or gained, but can be actively recalibrated through targeted conversational repair. Our results reveal that different strategies are variably effective depending on the domain, timing, and nature of the breakdown, which strongly argues against a universal repair mechanism. Instead, we advocate for the integration of adaptive, context-aware trust repair modules that can diagnose the breakdown type – be it factual, pragmatic, or tonal – and respond with a suitable repair strategy aligned with user expectations. For instance, while explicit corrections are critical in technical and fact-based exchanges, our data suggest that transparency and reflection are far more effective in complex or ethically nuanced dialogues, where users value the system’s acknowledgment of uncertainty or subjectivity. Furthermore, our trust trajectory modeling shows that speed of repair is critical: delayed or unacknowledged breakdowns often led to irreversible trust erosion, regardless of the eventual repair quality. This underscores the importance of proactive detection mechanisms embedded within the dialogue manager that can flag potential misalignments and initiate timely reparative action. The role of emotionally intelligent language – apologies, softeners, hedging, and metacognitive framing – also emerged as a subtle but powerful factor in trust recovery, especially among users who begin with lower baseline trust. Beyond the scope of this study, the modular nature of our trust recalibration framework makes it suitable for real-time deployment in a wide range of applications, from educational chatbots to customer service agents, and invites future work on learning-based strategy optimization, personalization based on user profiles, and cultural variations in trust perception. Ultimately, we argue that trust is not a precondition to conversational AI success – it is an outcome to be monitored, maintained, and repaired dynamically through thoughtful dialogue design.

V. CONCLUSIONS AND FUTURE WORK

This paper proposed a novel framework for modeling trust recalibration in conversational AI, positioning trust not as a static construct but as a dynamic state influenced by real-time dialogue behavior. Through simulated ChatGPT dialogues and annotated repair sequences, we introduced and validated a lightweight rule-based model that captures how user trust degrades and recovers across turn-level interactions. Our analysis demonstrated that trust can be actively managed through appropriate conversational repair strategies, and that effectiveness varies meaningfully across domains and strategy types. Importantly, our findings establish trust recovery as an actionable design objective for future AI dialogue systems.

Looking ahead, this work opens several promising research directions. Future studies should explore how the proposed trust recalibration model performs in real-world, longitudinal deployments involving diverse user populations. Adaptive strategy selection, informed by reinforcement learning or contextual modeling, could enhance the flexibility and personalization of trust repair. Additionally, extending the framework to account for multimodal cues, user profiles, and cultural trust norms will help create more robust, transparent,

and socially intelligent AI systems capable of sustaining long-term user engagement and confidence.

REFERENCES

- [1] A. Krouska, C. Troussas, I. Voyiatzis, P. Mylonas and C. Sgouropoulou, "ChatGPT-based Recommendations for Personalized Content Creation and Instructional Design with a Tailored Prompt Generator," *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*, Dubai, United Arab Emirates, 2024, pp. 295-299, doi: 10.1109/FLLM63129.2024.10852487.
- [2] F. Neha, D. Bhati, D. K. Shukla, and M. Amiruzzaman, "ChatGPT: Transforming healthcare with AI," *AI*, vol. 5, no. 4, pp. 2618–2650, 2024. [Online]. Available: <https://doi.org/10.3390/ai5040126>
- [3] C. Troussas, C. Papakostas, A. Krouska, P. Mylonas and C. Sgouropoulou, "Evaluating ChatGPT-driven Automated Test Generation for Personalized Programming Education," *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*, Dubai, United Arab Emirates, 2024, pp. 194-200, doi: 10.1109/FLLM63129.2024.10852510.
- [4] Ž. Bašić, A. Banovac, I. Kružić *et al.*, "ChatGPT-3.5 as writing assistance in students' essays," *Humanities and Social Sciences Communications*, vol. 10, p. 750, 2023. [Online]. Available: <https://doi.org/10.1057/s41599-023-02269-7>
- [5] C. Troussas, A. Krouska, P. Mylonas, C. Sgouropoulou, and I. Voyiatzis, "Fuzzy memory networks and contextual schemas: Enhancing ChatGPT responses in a personalized educational system," *Computers*, vol. 14, no. 3, p. 89, 2025. [Online]. Available: <https://doi.org/10.3390/computers14030089>
- [6] H. Su, W. Luo, Y. Mehdad, W. Han, E. Liu, W. Zhang, M. Zhao, and J. Zhang, "LLM-friendly knowledge representation for customer support," in *Proc. 31st Int. Conf. Comput. Linguistics: Industry Track*, Abu Dhabi, UAE, 2025, pp. 496–504.
- [7] C. Troussas, A. Krouska, C. Papakostas, P. Mylonas and C. Sgouropoulou, "Assessing the Impact of Integrating ChatGPT as an Advice Generator in Educational Software," *2024 9th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM)*, Athens, Greece, 2024, pp. 127-133, doi: 10.1109/SEEDA-CECNSM63478.2024.00031.
- [8] T. He and F. Jazizadeh, "Trust in human-AI interaction: Review of empirical research on trust in AI-powered smart home ecosystems," in *Computing in Civil Engineering 2023*, pp. 530–538. [Online]. Available: <https://doi.org/10.1061/9780784485224.064>
- [9] [32] H. Liu, Y. Wang, W. Fan, X. Liu, Y. Li, S. Jain, Y. Liu, A. Jain, and J. Tang, "Trustworthy AI: A computational perspective," *ACM Transactions on Intelligent Systems and Technology*, vol. 14, no. 1, Art. no. 4, Feb. 2023. [Online]. Available: <https://doi.org/10.1145/3546872>
- [10] V. L. Pop, A. Shrewsbury, and F. T. Durso, "Individual differences in the calibration of trust in automation," *Human Factors*, vol. 57, no. 4, pp. 545–556, 2015. [Online]. Available: <https://doi.org/10.1177/0018720814564422>
- [11] S. Shahrdrar, L. Menezes, and M. Nojournian, "A survey on trust in autonomous systems," in *Intelligent Computing. SAI 2018*, K. Arai, S. Kapoor, and R. Bhatia, Eds., *Advances in Intelligent Systems and Computing*, vol. 857. Cham, Switzerland: Springer, 2019. [Online]. Available: https://doi.org/10.1007/978-3-030-01177-2_27
- [12] P. K. Kahr, G. Rooks, M. C. Willemsen, and C. C. P. Snijders, "Understanding trust and reliance development in AI advice: Assessing model accuracy, model explanations, and experiences from previous interactions," *ACM Transactions on Interactive Intelligent Systems*, vol. 14, no. 4, Art. no. 29, Dec. 2024. [Online]. Available: <https://doi.org/10.1145/3686164>
- [13] J. Zerilli, U. Bhatt, and A. Weller, "How transparency modulates trust in artificial intelligence," *Patterns*, vol. 3, no. 4, p. 100455, 2022. [Online]. Available: <https://doi.org/10.1016/j.patter.2022.100455>
- [14] M. Poornikoo, W. Gyldesten, B. Vesin, and K. I. Øvergård, "Trust in automation (TiA): Simulation model, and empirical findings in supervisory control of maritime autonomous surface ships (MASS),"

- International Journal of Human–Computer Interaction*, pp. 1–28, 2024. [Online]. Available: <https://doi.org/10.1080/10447318.2024.2399439>
- [15] M. J. McGrath, O. Lack, J. Tisch, and A. Duenser, “Measuring trust in artificial intelligence: Validation of an established scale and its short form,” *Frontiers in Artificial Intelligence*, vol. 8, Art. no. 1582880, 2025. [Online]. Available: <https://doi.org/10.3389/frai.2025.1582880>
 - [16] Y. Yang and M.-Y. Kim, “Promoting sustainable transportation: How people trust and accept autonomous vehicles—Focusing on the different levels of collaboration between human drivers and artificial intelligence—An empirical study with partial least squares structural equation modeling and multi-group analysis,” *Sustainability*, vol. 17, no. 1, p. 125, 2025. [Online]. Available: <https://doi.org/10.3390/su17010125>
 - [17] Y. Li, B. Wu, Y. Huang, and S. Luan, “Developing trustworthy artificial intelligence: Insights from research on interpersonal, human-automation, and human-AI trust,” *Frontiers in Psychology*, vol. 15, Art. no. 1382693, 2024. [Online]. Available: <https://doi.org/10.3389/fpsyg.2024.1382693>
 - [18] J. Kauttonen, R. Rousi, and A. Alamäki, “Trust and acceptance challenges in the adoption of AI applications in health care: Quantitative survey analysis,” *Journal of Medical Internet Research*, vol. 27, p. e65567, 2025. [Online]. Available: <https://doi.org/10.2196/65567>
 - [19] E. Glikson and A. W. Woolley, “Human trust in artificial intelligence: Review of empirical research,” *Academy of Management Annals*, vol. 14, no. 2, pp. [online ahead of print], Aug. 2020. [Online]. Available: <https://doi.org/10.5465/annals.2018.0057>
 - [20] D. Zhang, T. T. Wijaya, Y. Wang *et al.*, “Exploring the relationship between AI literacy, AI trust, AI dependency, and 21st century skills in preservice mathematics teachers,” *Scientific Reports*, vol. 15, p. 14281, 2025. [Online]. Available: <https://doi.org/10.1038/s41598-025-99127-0>
 - [21] X. Zhao, W. You, Z. Zheng, S. Shi, Y. Lu, and L. Sun, “How do consumers trust and accept AI agents? An extended theoretical framework and empirical evidence,” *Behavioral Sciences*, vol. 15, no. 3, p. 337, 2025. [Online]. Available: <https://doi.org/10.3390/bs15030337>
 - [22] S. Mehrotra, C. Degachi, O. Vereschak, C. M. Jonker, and M. L. Tielman, “A systematic review on fostering appropriate trust in human-AI interaction: Trends, opportunities and challenges,” *ACM Journal of Responsible Computing*, vol. 1, no. 4, Art. no. 26, Dec. 2024. [Online]. Available: <https://doi.org/10.1145/3696449>
 - [23] A. Ferrario and M. Loi, “How explainability contributes to trust in AI,” in *Proc. 2022 ACM Conf. Fairness, Accountability, and Transparency (FAcT '22)*, New York, NY, USA, 2022, pp. 1457–1466. [Online]. Available: <https://doi.org/10.1145/3531146.3533202>
 - [24] P. Kahr, G. Rooks, C. Snijders, and M. C. Willemsen, “Good performance isn't enough to trust AI: Lessons from logistics experts on their long-term collaboration with an AI planning system,” in *Proc. 2025 CHI Conf. Human Factors in Computing Systems (CHI '25)*, New York, NY, USA, Art. no. 1100, pp. 1–16. [Online]. Available: <https://doi.org/10.1145/3706598.3713099>
 - [25] A. Klingbeil, C. Grützner, and P. Schreck, “Trust and reliance on AI—An experimental study on the extent and costs of overreliance on AI,” *Computers in Human Behavior*, vol. 160, p. 108352, 2024. [Online]. Available: <https://doi.org/10.1016/j.chb.2024.108352>
 - [26] M. Vössing, N. Kühl, M. Lind *et al.*, “Designing transparency for effective human-AI collaboration,” *Information Systems Frontiers*, vol. 24, pp. 877–895, 2022. [Online]. Available: <https://doi.org/10.1007/s10796-022-10284-3>