

# Trust Evaluation and Calibration in AI-infused Systems: The CRiDiT Computational Framework

The rise of artificial intelligence (AI), particularly generative AI, has intensified human-AI interaction, making the alignment of human trust to AI’s trustworthiness—known as trust calibration—a critical challenge. Misplaced trust, whether over- or under-trust leads to significant operational and safety risks. However, despite ongoing efforts toward trustworthy AI, there remains a lack of practical testbeds and insufficient attention to integrated risk evaluation and trust calibration. This paper instantiates CRiDiT (Computational Risk-Sensitive biDirectional Trust Model), a dynamic trust calibration framework synthesized from established computational trust processes. CRiDiT provides a structured blueprint for continuously monitoring human trust signals, machine performance, and contextual risk to adaptively calibrate trust across application domains. The framework maps calibration actions to specific trust factors within a given use case, enabling analysis of how those factors influence user trust. Following a Design Science Research (DSR) methodology, we conduct an observational case study (N=15) of CRiDiT through a prototype implementation across three realistic scenarios using behavioral metrics and qualitative feedback. The work contributes an operationalizable trust calibration artifact and establishes a foundation for future quantitative trust evaluations.

## ACM Reference Format:

. 2026. Trust Evaluation and Calibration in AI-infused Systems: The CRiDiT Computational Framework. 1, 1 (February 2026), 19 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

## 1 INTRODUCTION

Artificial intelligence systems are increasingly integrated into human life, creating complex human-AI interactions, from collaborative tools to autonomous agents. The effectiveness, reliability, and ethical use of these AI-infused systems depend not only on algorithmic accuracy but also on a complex and subjective human factor: trust.

Trust refers to a human’s belief regarding a system’s capacity to perform assigned tasks, with a willingness to accept uncertainty and vulnerability [17]. With the increasing integration of artificial intelligence (AI) into larger technical infrastructures—what we refer to as an “AI-infused system” [2]—the uncertainty of system behavior and output has become a great user concern, especially in high-stakes scenarios where system failure could lead to significant harm. For example, recently, **Open AI** has launched generative AI-based conversational systems for healthcare-related tasks. In such contexts, AI systems may directly influence user’s medical decision-making processes, where erroneous outputs, misinterpretations, or inappropriate reliance can lead to serious health consequences.

We adopt the Design Science Research (DSR) methodology [30], which is efficient for addressing concrete engineering problems through a rigorous cycle of artifact design and evaluation. We follow the DSR design cycle:

- **Problem Investigation.** We synthesize the observed core problem as “misplaced trust causing misassigned power”. Over-trust in a flawed system can lead to misuse, while under-trust can result in the underutilization of a capable tool, wasting potential resources and efficiency [19]. Within a broader shift toward integrating AI

---

Author’s address:

---

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

© 2026 Association for Computing Machinery.

Manuscript submitted to ACM

Manuscript submitted to ACM

into critical socio-technical infrastructures, ensuring appropriate alignment between human trust and system trustworthiness becomes a central human-AI interaction concern and a real engineering requirement for AI developers. This complex dynamic trust calibration process leads to two research questions:

- **RQ1:** Which trust factors have the strongest influence on human trust in the context of AI-infused systems?
- **RQ2:** Which trust remediation actions embedded in the system bring the strongest effect on human trust?

These two questions scope the “what” and “how” of the trust calibration process.

- **Treatment Design.** Recent literature reviews [7, 11, 12, 21, 27] have synthesized key components required for trust evaluation and calibration, including the measurement of subjective human trust, the objective evaluation of machine performance, contextual risk analysis, and adaptive remediation actions. Several trust modeling approaches have been presented, ranging from descriptive semantic frameworks to computational models that attempt to quantify trust. While descriptive models provide a strong conceptual foundation for understanding trust evaluation and calibration, their translation into operational engineering frameworks remains a challenge due to the lack of testbeds. Computational trust models aim to reduce this gap by enabling measurable and adaptive formalization of trust. Among computational trust approaches, the CRiDiT (Computational Risk-Sensitive biDirectional Trust) model is selected as an operational basis because it provides an explicit framework suitable for implementation in real-world scenarios and incorporates contextual risk sensitivity, aligning with the identified design requirements. Section 2 presents the relevant knowledge background and outlines the requirements.
- **Treatment Implementation and Observational Case Study.** To investigate the adaptability and the applicability of CRiDiT in operational settings, the model is implemented within a prototype testbed designed to observe trust evolution across distinct scenarios based on collected user behavior data and qualitative feedback. Human trust is evaluated through dynamic updates of model behavior based on interaction outcomes. This approach enables examination of CRiDiT’s operational feasibility as a computational trust calibration framework, while also uncovering patterns in how different trust factors and remediation actions influence human trust. Section 3 presents the concrete implementation of CRiDiT design, including its technical details. Section 4 describes the qualitative research methodology. Section 5 presents the results, analyzing how CRiDiT addresses the research questions, as well as the validation of the requirements. Finally, Section 6 concludes, discusses limitations, and outlines future work.

## 2 BACKGROUND

Trustworthy AI has become a research focus due to the increasing integration of AI systems into broader and safety-critical domains. Prior literature has investigated trust in AI systems through literature reviews and taxonomies of trustworthy AI and AI-related risks [1]. These efforts provide a consistent knowledge base on trust factors in the general context of AI, including regulatory, technical, social, and ethical dimensions.

While such structural and taxonomic approaches are valuable for defining trust and categorizing trust factors, they have limitations in supporting the measurement of trust and, consequently, in aligning human trust perception with system reliability. The concept of computational trust, introduced by Marsh in 1994 [18], aimed to reduce this gap by moving from structural trust definitions toward operational trust evaluation. Computational trust approaches provide an operationalizable way to quantify trust and guidance on how trust factors can be computationally integrated and updated over time to support dynamic trust calibration during interaction.

In parallel, prior work on computational trust has proposed mathematical models for representing trust either as a subjective psychological state using probabilistic approaches, or through aggregation of weighted trust factors and

contextual impacts, or through information fusion techniques for handling trust-related evidence. These approaches offer mechanisms to quantify trust and support dynamic updating [4, 10, 13, 20]. Behavioral evaluation approaches emphasize experimentally grounded measures of trust level through behavioral metrics, such as BLOCKIES [14]. Together, these approaches provide prerequisites for trust calibration in interactive AI systems.

Recent research has proposed several approaches to trust calibration. Learning-based calibration mechanisms anchor calibration decisions in observed system outcomes, for example, Yu et al.[32] illustrate how trust evolves in relation to system performance. Predictive models consider human trust as a temporal process and aim to anticipate future human behavior in order to identify potential miscalibration during human-AI interaction, such as Wang et al.’s work [28]. Other frameworks focus on the role of explainability and confidence cues in influencing human trust, including AXTF (Adaptive Explainability Trust Framework) [9], and Yu et al.’s computational framework [33]. These approaches typically address individual dimensions of trust calibration—such as explainability, performance evaluation, or trust prediction—in isolation. By contrast, conceptual calibration frameworks such as TCMM (Trust Calibration Maturity Model) [26], IMPACTS (Intention, Measurability, Performance, Adaptivity, Communication, Transparency, Security) [12] and CRiDiT (Computational Risk-Sensitive biDirectional Trust) structure trust factors, risks, and (re)calibration processes at a system level, providing guidance for analysis and evaluation.

While prior work establishes important foundations for trust calibration, translating these concepts into an operational AI-infused system introduces additional engineering constraints. CRiDiT is selected as an operational basis because it is an operational blueprint that explicitly connects trust inputs, computational trust functions, and system-embedded remediation actions, making it suitable as an operational basis for implementation in AI-infused systems.

Operationalization introduces concrete engineering implications. First, trust evaluation and calibration must explicitly account for contextual risk. Second, the framework must capture the dynamic nature of trust. Third, computational trust representation must be measurable and comparable across human and system perspectives in order to support calibration decisions. Finally, calibration decisions must be mapped to concrete remediation actions, such as improved explainability and transparency, or system behavior updates. Therefore, Table 1 presents the requirements ensuring the resulting artifact remains (1) risk-sensitive, (2) dynamically manages trust, and (3) executes concrete remediation actions for trust calibration.

Table 1. Goal and Requirements for Trust Calibration Framework Implementation

<b>Design Goals</b>	<b>Design Requirements</b>	
Risk Sensitivity	R1	Perceived risk must be collected, it directly influences human trust computation.
	R2	Remediation actions must be triggered based on contextual risk level, with higher risk demanding faster response.
Dynamic Trust Management	R3	Trust scores and calibration status must be updated in real-time based on interactions.
	R4	Subjective human trust signals must be quantified to be comparable with machine trust scores.
	R5	Technical evidence must be quantified and fused for machine trust computation, starting from a benchmark performance baseline.
Remediation Actions	R6	Calibration status (over-trust, under-trust, well-calibrated) must map to concrete remediation actions.

The listing design requirements (Table 1) serve as implementation criteria for the operational realization of CRiDiT described in Section 3.

### 3 CRIDIT IMPLEMENTATION

The CRiDiT aims to address the three main goals as listed in Tab. 1. The design respect the need of design requirements. In this section, we present the concrete implementation practices for each layer of CRiDiT.

#### 3.1 CRiDiT Architecture

**3.1.1 Overview.** The subsequent subsections detail the concrete implementations illustrated in Figure1 that illustrates the overview of the conceptual trust model CRiDiT. Table2 illustrates the mathematical notations we assigned to different notions.

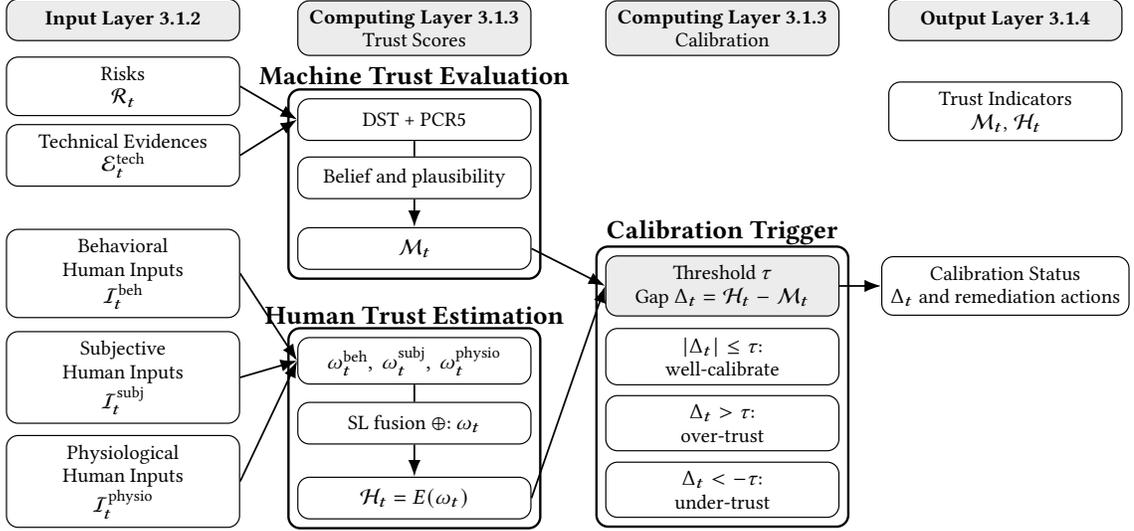


Fig. 1. CRiDiT Overview

**3.1.2 Input Layer.** This layer collects trust factors across three dimensions: socio-ethical considerations, technical characteristics, and human inputs [3]. We further categorize these trust factors as follows:

- **System-Related Trust Factors:** these factors are derived from objective evidence that enables humans to assign a trust level through cognitive judgment. Such evidence generally follows technical benchmarks specific to the application context. Positive evidence includes system reliability metrics, negative evidence includes system failures, and uncertainty-related evidence includes risk events.
- **Human-Related Trust Factors:** these factors encompass both external influences, such as knowledge of the system provider’s reputation, perception of system’s ethical behavior, and prior experience with similar systems, as well as internal variabilities, such as dispositional personality traits, gender, or age. These internal factors are independent of the specific system in use.

Measuring human inputs is challenging due to individual subjectivity and the ambiguous representativeness of certain metrics (e.g., heart rate and stress level). In this work, we consider the adoption of system outputs as a direct observation of trust behavior. User feedback provides an indirect indication of a user’s trust level,

Table 2. CRiDiT Notations

Notations	Description
$X = \{T, UT\}$	The set of all possible states: trustworthy T or untrustworthy UT.
$2^X = \{\emptyset, \{T\}, \{UT\}, \{T, UT\}\}$	The power set that includes all the possible combinations of elements in the set $X$ .
$\mathcal{T}_t, \mathcal{N}_t$	Threats and negative impacts.
$\mathcal{V}_t^{\text{sys}}, \mathcal{V}_t^{\text{model}}, \mathcal{V}_t^{\text{hcom}}$	System, model, and human communication level vulnerabilities.
$\mathcal{R}_t = \{\mathcal{T}_t, \mathcal{N}_t, \mathcal{V}_t^{\text{sys}}, \mathcal{V}_t^{\text{model}}, \mathcal{V}_t^{\text{hcom}}\}$	Perceived risks at time $t$ .
$\mathcal{E}_t^{\text{tech+}}, \mathcal{E}_t^{\text{tech-}},$ $\mathcal{E}_t^{\text{tech}} = \mathcal{E}_t^{\text{tech+}} \cup \mathcal{E}_t^{\text{tech-}}$	Positive, negative and combined technical evidence sets.
$\mathcal{E}_t = \mathcal{E}_t^{\text{tech}} \cup \mathcal{R}_t$	The union set of technical evidence and perceived risks.
$N_{\mathcal{E}_t}$	The number of evidence collected (trust factors and perceived risks) at time $t$ .
$e_i, X_{e_i}, m_{e_i}$	The $i^{\text{th}}$ evidence in $\mathcal{E}_t$ , its state, and its belief mass.
$X_{\mathcal{E}_t}$	The set of all the $X_{e_i}$ .
$m_t^{\text{cons}}(x)$	The belief mass of state $x$ resulting from consensus fusion.
$m_t^{\text{conf}}$	The belief mass resulting from conflict fusion.
$m_{\mathcal{E}_t}(x), x \in 2^X \setminus \{\emptyset\}$	The global belief mass on state $x$ .
$bel_t(\{T\})$	The belief of trust state at time $t$ .
$pl_t(\{T\})$	The plausibility of trust state at time $t$ .
$\mathcal{M}_t$	The system trustworthiness score.
$\mathcal{I}_t^{\text{beh}}, \mathcal{I}_t^{\text{subj}}, \mathcal{I}_t^{\text{physio}}$	Behavioral, subjective, and physiological human trust inputs.
$b_t^T, d_t^T, u_t^T, a^T$	Belief, disbelief, and uncertainty in $\{T\}$ , and base rate, at time $t$ .
$\omega_t^{\text{beh}}, \omega_t^{\text{subj}}, \omega_t^{\text{physio}}$	Opinions based on behavioral, survey and physiological inputs.
$\omega_t$	Fused opinion based on behavioral, survey and physiological opinions.
$\mathcal{H}_t$	The estimated human trust score.

for example, the statement “I drank the medicine proposed by the chatbot, and now I feel worse” reflects a perceived system malfunction and a likely decrease in trust, although such effects remain dependent on individual characteristics. Physiological signals constitute the least direct indicators, for instance, prolonged eye fixation does not necessarily imply high attentional capacity.

In addition to these trust factors, Lee and See’s [17] definition of trust emphasizes its strong relationship with risk. From previous work, we indicated that risk analysis must incorporate the threats [31], their potential negative impacts on human [29], and vulnerabilities at three distinct levels: (1) system-level, (2) model-level, and (3) human-communication-level.

**3.1.3 Computing Layer.** This layer defines the computation of measurements from the inputs layer using appropriate trust metrics. We treat machine-side trust evaluation and human trust perception separately, then compare their values to identify potential over-trust or under-trust. Finally, we define trust calibration mechanisms to address such misplaced trust. The subsequent subsection discusses the outputs designed to enable user-side trust calibration.

*Machine-side Trust Evaluation  $\mathcal{M}_t$ .* There are two dimensions for assessing trust on machine-side evaluation: (1) fusion of technical design evidence from benchmarks, and (2) risk assessment.

The relation between trust factors is not linear, two trust factors can be in conflict and cause problems when fusing the measurement for determining if the system is trustworthy or not [8], plus, they are strongly related to risks on different levels, so risk assessment not only introduces negative impacts on trust, these paradoxes further exacerbate the challenge. The Dempster-Shafer Theory (DST) is a powerful tool for handling such complex data fusion [4].

We adopt the mathematical notations in Table 2. Note that each element in the power set represents a belief about the system,  $\emptyset$  is the conflicting state between evidences, Trustworthy is the complete positive trust level, Untrustworthy is the complete negative trust level, and  $\{\text{Trustworthy}, \text{Untrustworthy}\}$  represents the uncertainty. For example, we have  $m_{\text{accuracy}}(\{\text{Trustworthy}\}) = 0.8$  which means with the measurement and the benchmark of the AI-infused system in question, the system is highly trustworthy,  $m_{\text{accuracy}}(\{\text{Untrustworthy}\}) = 0$  which means there is no direct evidence proof that with such measurement the system is untrustworthy, and  $m_{\text{accuracy}}(\{\text{Trustworthy}, \text{Untrustworthy}\}) = 0.2$  which means because of the accuracy is not 100%, it includes the possibility that “the inaccuracy caused distrust”.

For the global mass calculation, the DST model distinguishes two parts: (1) the consensus conjunctive, where the joint of features’ states is the state in evaluation, and (2) the conflicting mass, where the joint is an empty set [22, 23]. Initially, the consensus mass and the conflict mass can be obtained by the conjunctive combination of evidence as in classical DST. Then, the conflict redistribution can be handled by normalization using Dempster’s rule, however the limitation of this rule is that it may discard strong negative evidence among massive positive evidence. Therefore, to avoid this limitation, we adapt the Proportional Conflict Redistribution rule n<sup>5</sup> (PCR5), which redistributes the conflicting mass proportionally [5, 23, 24].

Formally, for a set containing  $N_{\mathcal{E}_t}$  pieces of evidence,  $\forall x \in 2^X \setminus \{\emptyset\}$ ,  $\forall X_{\mathcal{E}_t} \subseteq 2^X$ , the belief mass for the consensus conjunctive is expressed as:

$$m_t^{\text{cons}}(x) = \sum_{X_{\mathcal{E}_t}} \prod_{i=1}^{N_{\mathcal{E}_t}} m_{e_i}(X_{e_i}) \quad (1)$$

$$\bigcap_{i=1}^{N_{\mathcal{E}_t}} X_{e_i} = x$$

and the belief mass for conflict conjunctive is expressed as:

$$m_t^{\text{conf}} = \sum_{X_{\mathcal{E}_t}} \prod_{i=1}^{N_{\mathcal{E}_t}} m_{e_i}(X_{e_i}) \quad (2)$$

$$\bigcap_{i=1}^{N_{\mathcal{E}_t}} X_{e_i} = \emptyset$$

Within the PCR5 redistribution, we have the belief mass expressed as:

$$m_{\mathcal{E}_t}(x) = m_t^{\text{cons}}(x) + \sum_{X_{\mathcal{E}_t}} \left( \frac{m_{e_i}(x)^2 \prod_{\substack{k=1 \\ k \neq i}}^{N_{\mathcal{E}_t}} m_{e_k}(X_{e_k})}{m_{e_i}(x) + \sum_{\substack{k=1 \\ k \neq i}}^{N_{\mathcal{E}_t}} m_{e_k}(X_{e_k})} \right) \quad (3)$$

$$\bigcap_{i=1}^{N_{\mathcal{E}_t}} X_{e_i} = \emptyset$$

$$X_{e_i} = x$$

We obtain a scalar trust estimate from DST by the belief  $bel_t(\{T\})$  and plausibility  $pl_t(\{T\})$  bounds.

The belief of the trust state being valid is the mass of its only proper subset, itself, we have:

$$bel_t(\{T\}) = m_{\mathcal{E}_t}(\{T\}) \quad (4)$$

And the plausibility is the sum of all the masses of the other subset in the power set whose intersection with  $\{T\}$  is not  $\emptyset$ .

$$pl_t(\{T\}) = m_{\mathcal{E}_t}(\{T\}) + m_{\mathcal{E}_t}(\{T, UT\}) \quad (5)$$

To enable the machine to make a hard decision on system's trust state, we use the Pignistic transform [25], we have:

$$\text{BetP}_t(\{T\}) = m_{\mathcal{E}_t}(\{T\}) + \frac{m_{\mathcal{E}_t}(\{T, UT\})}{2}$$

With (4) and (5), we observe that for the binary atomic state set  $X = \{T, UT\}$ , the Pignistic transform equals the midpoint between belief and plausibility:

$$\text{BetP}_t(\{T\}) = \frac{bel_t(\{T\}) + pl_t(\{T\})}{2}$$

We define the machine-side trust estimate as this Pignistic probability:

$$\mathcal{M}_t = \text{BetP}_t(\{T\}) \quad (6)$$

*Human Trust Perception.* Unlike the machine-side trust evaluation, calculating the human trust level has fewer directly observable measurements. We mentioned three types of human inputs: behavioral (if human accepted the output of the AI-infused system), subjective (the feedback of users through surveys or ratings), and physiological signals (helping analyze human psychological states). We adapt the Subjective Logic model of Jøsang [15, 16] for handling these subjective opinions, which explicitly represents belief, disbelief, uncertainty, and base rates.

We use the Subjective Logic notations introduced in Table 2. For behavioral human inputs, observations are binary (accepted or rejected), therefore, we use Beta binomial distribution to model the associated opinion [16]. Let  $p_t^{T,beh} = p_t^{beh}(\{T\})$  the probability of human acceptance,  $r_t^{T,beh} = r_t^{beh}(\{T\})$  the total number of acceptance and  $s_t^{T,beh} = s_t^{beh}(\{T\})$  the total number of rejections. From the historical behavioral observations, we infer a predictive distribution for trust-related behavior. The probability density of  $p_t^{T,beh}$  followed Beta distribution:

$$\text{Beta}(p_t^{T,beh}, \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} (p_t^{T,beh})^{\alpha-1} (1 - p_t^{T,beh})^{\beta-1}$$

where  $\alpha, \beta > 0$  are given by:

$$\begin{cases} \alpha = r_t^{T,beh} + a^{T,beh}W^{beh} \\ \beta = s_t^{T,beh} + (1 - a^{T,beh})W^{beh} \end{cases}$$

Here  $a^{T,beh}$  is the base rate for behavioral trust before any interactions and  $W^{beh} = 2$  is the weight for binomial opinions.

From this distribution, we derive the corresponding Subjective Logic opinion on the proposition “the system is trustworthy” based on behavioral inputs. This opinion is expressed with the belief derived from acceptance  $b_t^{T,beh}$ , disbelief  $d_t^{T,beh}$  derived from rejections, uncertainty  $u_t^{T,beh}$  due to limited observations:

$$\omega_t^{beh} = (b_t^{T,beh}, d_t^{T,beh}, u_t^{T,beh}, a^{T,beh})$$

Where

$$\begin{cases} b_t^{T,beh} = \frac{r_t^{T,beh}}{W^{beh} + r_t^{T,beh} + s_t^{T,beh}} \\ d_t^{T,beh} = \frac{s_t^{T,beh}}{W^{beh} + r_t^{T,beh} + s_t^{T,beh}} \\ u_t^{T,beh} = \frac{W^{beh}}{W^{beh} + r_t^{T,beh} + s_t^{T,beh}} \end{cases}$$

In terms of verbal feedback from surveys or ratings, we will use the qualitative opinion representation [16], Jøsang put likelihood levels and confidence levels into a matrix. We map elements in the matrix (likelihood x confidence) to a point in the opinion triangle  $(b, d, u)$  using  $b = c\ell$ ,  $d = c(1 - \ell)$ , and  $u = 1 - c$ , where  $\ell$  is likelihood and  $c$  is confidence. We denote  $\omega_t^{\text{subj}}$  as the human perception of trustworthiness based on surveys and  $\omega_t^{\text{physio}}$  as the opinion from physiological signals.

We fuse these opinions with hierarchical weighting that prioritizes behavioral inputs (most direct) over subjective inputs (self-reported), and subjective inputs over physiological signals (most indirect). Let  $w_i$  be input weights for behavior, subjective, and physiological channels. We apply a power transform  $w_i^\gamma$  (with  $\gamma = 2$ ) and normalize to emphasize dominant channels. Let  $\tilde{w}_i = w_i^\gamma / \sum_j w_j^\gamma$  denote the normalized transformed weights; these  $\tilde{w}_i$  are the  $w^{\text{beh}}$ ,  $w^{\text{subj}}$ ,  $w^{\text{physio}}$  used below. Each opinion is discounted by its weight and then combined with consensus:

$$\begin{aligned}\omega_t^{\text{beh}'} &= \tilde{w}^{\text{beh}} \odot \omega_t^{\text{beh}}, \\ \omega_t^{\text{subj}'} &= \tilde{w}^{\text{subj}} \odot \omega_t^{\text{subj}}, \\ \omega_t^{\text{physio}'} &= \tilde{w}^{\text{physio}} \odot \omega_t^{\text{physio}}, \\ \omega_t &= \omega_t^{\text{beh}'} \oplus \omega_t^{\text{subj}'} \oplus \omega_t^{\text{physio}'}\end{aligned}$$

Here,  $\odot$  denotes weight-based trust discounting and  $\oplus$  denotes the consensus fusion operator, both defined in Subjective Logic.

With  $\omega_t^{\mathcal{H}} = (b_t^T, d_t^T, u_t^T, a^T)$  as the human trust perception at time  $t$ , the trust score from human perception is the expected probability of trustworthiness from the fused human inputs:

$$\mathcal{H}_t = E(\omega_t) = b_t^T + a^T u_t^T$$

*Trust Calibration.* We define the gap  $\Delta_t = \mathcal{H}_t - \mathcal{M}_t$  and a threshold  $\tau$  as follows:

- $|\Delta_t| \leq \tau$ : the trust level can be considered as well-calibrated.
- $\Delta_t > \tau$ : the human over-trusts the system.
- $\Delta_t < -\tau$ : the human under-trusts the system.

The definition of threshold depends on the risk analysis, the greater the risk level is, the smaller the threshold should be to increase the sensitivity of captured failures.

**3.1.4 Output Layer.** In the output layer, we communicate two categories of trust-related inferences to the human user: (1) the model-estimated trust scores:  $\mathcal{M}_t$ , as well as the human trust score  $\mathcal{H}_t$ , and (2) the calibration status (under-trust or over-trust) along with the value of the gap and triggered trust remediation actions.

The trust evaluation process is rerun using updated system failure observations and human inputs, allowing refine us to trust estimations and reduction of the gap between system reliability and perceived trustworthiness. We monitor the evolution of system behavior and human adoption, subjective subscales and perceived transparency and the usefulness of trust cues as well as the effectiveness of machine-side behavior adaptation mechanisms.

### 3.2 Prototype Implementation

We instantiate the CRiDiT architecture within a chatbot prototype to operationalize the design requirements (R1-R6). The instantiation targets three distinct real-world high-stakes scenarios: (1) corporate financial scenario, (2) legal

scenario, and (3) hiring scenario. CRiDiT is implemented as a lightweight prototype that integrates with the OpenAI API under token constraints

3.2.1 *Operationalization of CRiDiT Constructs.* The data collection feeds directly into the computational layers of CRiDiT as implemented in the prototype.

*Machine-side Trust Evaluation.* Evidence set  $\mathcal{E}_t$ :

- **Negative Evidence:** observers can inject prompts for incomplete answers or send pre-scripted flawed answers containing trust violated answers, for instance, in the legal scenario, the response contains bias toward one position. These are assigned a mass value for the untrustworthy event  $\{UT\}$  and uncertainty event  $\{T, UT\}$ .
- **Positive Evidence:** While observers can inject prompt in a negative way, they can also inject prompt to test the remediation action, such as “add more explanations on your logic flow” or “show me exactly your uncertainty and your limits”. These are assigned a mass for the trustworthy event  $\{T\}$ .
- **Uncertainty Evidence:** Adapting the ISO definition, risk is the effect of uncertainty. While we inject negative prompt, as well as positive prompt, the response and the prompts themselves includes uncertainty, for example, while the prompt itself is incomplete and the prompt is “only show three most critical risk”, then the incompleteness in response we are not certain if it is caused by the system’s performance or the incompleteness in the source data. This is uncertainty evidence, assigned a mass for the uncertainty event  $\{T, UT\}$ , which impacts the system’s trustworthiness. For instance, this category includes prompt injection flaws (intentional threats), misconfigurations (unintentional threats), exposed AI algorithms (vulnerabilities), misinterpretation (negative impacts).

The initial machine trust score  $\mathcal{M}_0$  is the baseline performance using LLM benchmarks (MMLU). This benchmark is chosen for its generative adaptation as an initial proxy for general LLM capability. In this observational study,  $\mathcal{M}_0$  serves primarily as a starting point for demonstrating the update mechanisms. The ongoing machine trust score  $\mathcal{M}_t$  is updated using the positive and negative prompt events injected by observers. Each event is mapped by: (1) base masses from polarity: positive events assign mass to  $\{T\}$ , negative events assign mass to  $\{UT\}$  and  $\{T, UT\}$ ; and (2) severity scaling: the magnitude of deviation from expected behavior scales the mass (e.g., minor inaccuracy 0.2, major error 0.8).

*Human Trust Estimation.*

- Behavioral Human Trust Inputs: These are derived from interactions with AI systems.
- Subjective Human Trust Inputs: These consist mainly of self-reported rating through a communication interface, evaluating different trust factors, we collect self-confidence, reliability, predictability, task criticality, and task complexity. Additionally, we ask participants to think aloud, so the observers can write their remarks in real-time.
- Physiological Human Trust Inputs: While physiological signals are part of the CRiDiT model, they may lack representativeness in a chatbot use case, for instance, the eye fix on the screen do not represent that they are more concentrate, therefore these measurements are not incorporated into the implementation.

The initial human trust score  $\mathcal{H}_0$  is derived from the pre-flight questionnaire. We interpret  $\mathcal{H}_0$  as the expected trustworthiness  $\mathcal{H}_0 = E(\omega_0)$ , an aggregated normalized score from the Likert scales. To obtain a corresponding Subjective Logic opinion  $\omega_0 = (b_0^T, d_0^T, u_0^T, a^T)$ , we use a fuzzy-logic mapping: triangular membership over  $\mathcal{H}_0$  (low, medium,

high) interpolates the base rate  $a^T$  and uncertainty  $u_0^T$ ; we then compute belief  $b_0^T = \max(0, \min(1 - u_0^T, \mathcal{H}_0 - a^T * u_0^T))$  and disbelief  $d_0^T = 1 - b_0^T - u_0^T$ . During the interaction phase, the ongoing human trust score is updated from behavior inputs (acceptance/rejection of system outputs) and self-reported trust indicators. The behavioral inputs are modeled as a binomial opinion with prior weight ( $W = 2$ ). Self-reported indicators—self-confidence, reliability, predictability, task criticality, and task complexity—, are transformed into a Subjective Logic opinion via the qualitative-matrix mapping [16]: likelihood is the average of reliability and predictability ratings, and confidence is taken from self-confidence, and the base rate is adjusted by task criticality. The behavioral and subjective opinions are fused with weighted consensus (behavior weight 0.7 and subjective weight 0.3) to produce the human trust score  $H_t$ .

*Calibration.* Trust calibration is triggered when the absolute gap  $|\Delta_t| = |\mathcal{H}_t - \mathcal{M}_t|$  exceeds a fixed risk-sensitive threshold  $\tau$ . The threshold is set lower for higher-risk scenarios to increase sensitivity to miscalibration: we define 0.06 for the legal scenario, 0.07 for the financial scenario and 0.08 for the hiring scenario. Threshold values were selected through piloting to generate observable calibration events within session time constraints. They operationalize the concept of risk-sensitive triggering for demonstration purposes.

Initial trust cues presented to participants include the initial trust scores ( $\mathcal{M}_0, \mathcal{H}_0$ ), the initial trust calibration status, and warnings if over-trust or under-trust is detected. For each interaction, the gap  $\Delta_t$  is recomputed. If  $|\Delta_t| = |\mathcal{H}_t - \mathcal{M}_t|$ , the calibration status is updated and remediation actions are triggered. In the case of under-trust, the message “The system performs well in general. You can rely on it to improve efficiency.” is shown on the CRiDiT communication interface, along with corrective prompts injected by observers; for over-trust, “Please verify the output carefully. The system can make mistakes.” is shown, along with explanatory cues; for well-calibrated, no action is triggered.

**3.2.2 Interaction Flow.** The prototype instantiates the core CRiDiT components at a conceptual and interface level: (1) a simulated AI backend that evaluates and generates trust calibration mechanisms, (2) a communication interface that collects participant interaction data and presents trust cues to the participant, and (3) a management interface for observers, where the chatbot updates its behavior through prompts generated by observers in the CRiDiT calibration block. observers also in charge of reporting the ongoing human trust inputs to the backend.

Figure 2 is a communication diagram that illustrates the workflow using CRiDiT in the case of chatbot: (1) pre-flight questionnaire initializes  $\mathcal{H}_0$ , and observers set  $\mathcal{M}_0$  and  $\tau$ , (2) initial trust scores are displayed, (3) participants interact, observers inject prompts, and CRiDiT updates  $\mathcal{M}_t$ , (4) participants provide feedback (accept/reject, self-reports, think-aloud), and CRiDiT updates  $\mathcal{H}_t$ , (5) CRiDiT recomputes the gap and triggers remediation actions if thresholds are exceeded, (6) post-flight questionnaire closes the loop.

## 4 EVALUATION METHODOLOGY: OBSERVATIONAL CASE STUDY

We adopt an observational case study design with embedded mixed-methods data collection. This approach maintains distance from participants, allowing the collection of participants’ natural trust evolution without direct intervention. This qualitative study enables studying complex socio-technical phenomena: trust within real-world contexts. It seeks to understand how the trust factors impact participants’ trust, thereby addressing our design requirements (R1-R6). The case study focuses on cross-case characteristics of trust calibration, enabling identifying common patterns and context-specific factors in this complex process.

The data collection methodology integrates: (1) pre- and post-flight questionnaires for subjective and dispositional measures, and (2) an observational protocol for real-time behavioral and interaction data.

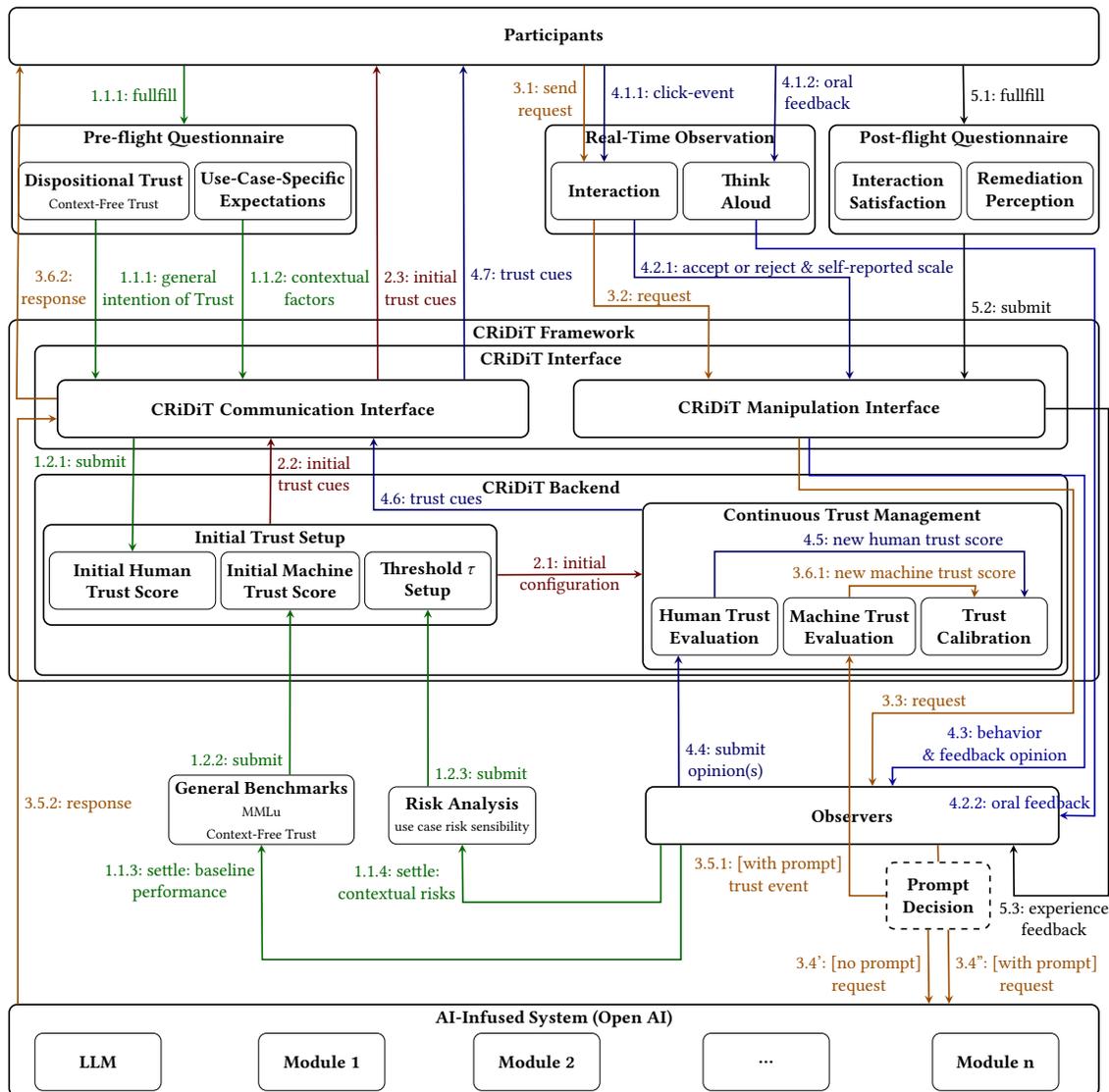


Fig. 2. CRiDiT Prototype Communication Diagram in Chatbot Use Case

#### 4.1 Data Collection

We collect data through two streams: questionnaires for subjective data and observational protocols for behavioral data. All the self-reported scaling questions in questionnaires or during the interaction phase are inspired by the Human-Computer Trust (HCT) questionnaire [6]. We inspect the personal intention and aversion, perceived system reliability, and perceived system's intention. We divided the questionnaire into three categories: (1) dispositional trust, which is context-free, scaling human general intention of trust in AI-infused system, include their expectation level and risk aversion level; (2) domain- and task-specific trust, which test human trusting beliefs and trusting intention on specific

domain and even more specifically on precise tasks; and (3) synthesizing questions which answered by participant after reflection and memory on past interactions. The following subsections detail the trust categories that each data collection phase is in charge of.

*4.1.1 Pre- and Post-Flight Questionnaire.* The questionnaires enable the assessment of trust patterns and correlations between perceived trustworthiness and individual differences across samples. We administer group-administered survey to participants (target  $N = 15$ ) before and after the interaction phase with the CRiDiT prototype:

Participants complete a pre-flight questionnaire assessing baseline states and contextualize interaction expectations:

- dispositional trust in AI technology (including their trust intention, the risk aversion, and the privacy concern);
- their familiarity with the domain;
- opinion on trust factors that are potentially important for the scenario;
- Two open questions on what break their trust and what will foster their trust in the chatbot use case.

The pre-flight questionnaire will generate the initial human trust score.

After the interaction phase, participants complete a post-flight questionnaire evaluating the interaction outcome and CRiDiT’s perceived impact:

- post-intervention trust in the chatbot;
- willingness to continue using the chatbot;
- perceived performance on trust factors;
- perceived changes in system with remediation actions;
- Two open questions on their perception on trust evolution during the interaction phase and their expectation on future adoption of chatbot in the domain.

Except the open questions, all other questions are on a 1-7 Likert scale.

*4.1.2 Observational Protocol.* Unlike questionnaires, the observational protocol allows us to collect real-time qualitative behavioral data during the interaction phase with the CRiDiT prototype.

For each decision point or system output within a scenario, participants are presented with trust cues as references, we then record the following:

- CRiDiT-generated outputs: machine trustworthiness score  $\mathcal{M}_t$ , human trust estimate  $\mathcal{H}_t$ , calibration gap  $\Delta_t$ , calibration status, and potential warnings depending on the calibration status;
- self-reported rating on self-confidence (1-7 likert scale);
- self-reported rating on system’s behavior (1-7 likert scale);
- self-reported rating on task nature perception (1-7 likert scale);
- real-time: any potential oral feedback from participants (think-aloud);
- behavioral choice: accept, reject, or further queries followed by the system’s responses.

## 4.2 Participants and Procedure

We targeted 15 participants, each participant will experience one primary scenario in a within-subjects design. Five participants for each scenario, 3 of them have relevant background on the scenario (graduated from the domain or working in the domain) and 2 of them have no experiences on the domain.

The procedure for a session is as follows:

- (1) pre-flight questionnaire;
- (2) tutorial on the prototype interface;
- (3) interaction phase: a series of two or three tasks within the assigned scenario. Tasks are classical automation tasks, including information retrieving, classification, text generation. Participants are given artificial inputs. The CRiDiT framework runs silently, observers both inject failures and enacted remediations without telling participants. Trust cues (trust scores, warnings while over-trust or under-trust detected) are displayed to participants via the interface.
- (4) post-flight questionnaire.

### 4.3 Abductive Inference Design

For an observational case study, besides the statistics that we can collect, such as acceptance rate or the self-reported scales, think-aloud feedback and trust-related behavior are difficult to capture through quantitative metrics alone. Despite the statistical analysis on quantified trust score evolution and the action rate reflecting trust patterns, we need to bring further explanations on the phenomena and the observations that we found during the execution phase.

Therefore, we adapt the abductive inference design, aiming to bring explanations on the observations, especially when unexpected issues occur (e.g., task misunderstandings, unexpected human behavior, etc). The analysis is structured into three steps: (1) we collect the descriptive summaries from raw interaction data (participant feedback and explanations on their behaviors); (2) we generate explanations for observed patterns in trust evolution, grounded in the CRiDiT interaction context (including prompts and other remediation actions); and finally (3) we formulate generalizations that specify the scope in which these explanations plausibly hold. All findings are interpreted as hypothesis-generating.

## 5 RESULTS

### 5.1 Descriptive Statistics

Table.3 summarizes the pre-flight questionnaire, interaction phase, and post-flight questionnaire results.

The average trust intention score was 5.021/7, indicating a general tendency to trust the chatbot. However, due to the high performance baseline score (MMLU of GPT-5 = 0.925), all participants started in an under-trust state. Participants in the legal scenario showed the lowest trust intention (4.67/7), followed by financial (5.13/7) and hiring (5.22/7). This ordering aligns with our hypothesis that higher-stakes domains elicit lower initial trust.

The average privacy concern score was 5.338/7, indicating strong concern about personal data security (e.g., “I am afraid that the more I interact with the chatbot, the better it knows me; I don’t know what they (the technical providers) are going to do with that data.”). The average risk aversion score was 4.728/7. Experts tended to spend more time reading and verifying LLM outputs, likely due to greater awareness of risks and consequences. Experts averaged 16.5 interaction iterations, compared to 8.0 for non-experts; non-experts tended to follow task descriptions without further querying, suggesting greater warning efforts to anchor verification.

Risk aversion is highly personal, so calibration-status counts provide a clearer view of trust evolution. The legal scenario maintained the lowest trust level, consistent with lower perceived system performance. The 21 well-calibrated status, combined with think-aloud feedback, show that while citation inaccuracies are unacceptable and directly reduce human trust, incompleteness can be tolerated and adapted to. One participant explained: “My expectations for AI differ by task. For the first legal task, I expected accurate citations, so incompleteness and inaccuracy were unacceptable. For

Table 3. Summary statistics from preflight, postflight, and calibration history.

Scenario	Pre-flight Questionnaire	Interaction Phase	Post-flight Questionnaire
Hiring	Trust Intention: 5.22 Risk Aversion: 5.72 Privacy Concern: 5.83 Fairness Importance: 6.33 Accuracy Importance: 5.83 Explainability Importance: 5.17 Transparency Importance: 5.00	Under-Trust: 14 Over-Trust: 19 Well-Calibrated: 16	Transparency: 5.17 Understandability: 5.00 Observed Fairness: 5.17 Appropriate Tone: 5.67 Willingness to Reuse: 5.33 Remediation (Fixed Problem): 4.50 Effective Remediation: Explanation: 3, Correction: 2 Remediation (Explanation Helpfulness): 3.17 Remediation (Trust Increment): 4.50
Financial	Trust Intention: 5.13 Risk Aversion: 4.33 Privacy Concern: 4.96 Source Grounding Importance: 6.20 Transparency Importance: 5.80 Consistency Importance: 5.80 Accuracy Importance: 5.80 Predictability Importance: 5.50	Under-Trust: 12 Over-Trust: 21 Well-Calibrated: 12	Calculation Accuracy: 5.50 Transparency (Communicated Uncertainty): 4.33 Traceability (Logic Flow): 5.00 Consistency Maintained: 5.50 Risk Completeness: 5.00 Willingness to Reuse: 5.00 Remediation (Fixed Problem): 4.33 Effective Remediation: Explanation: 1, Correction: 4 Remediation (Explanation Helpfulness): 4.17 Remediation (Trust Increment): 4.17
Legal	Trust Intention: 4.67 Risk Aversion: 3.93 Privacy Concern: 5.12 Reliability Importance: 6.40 Citation Accuracy Importance: 6.00 Accountability Importance: 5.00 Privacy Security Importance: 5.00	Under-Trust: 22 Over-Trust: 9 Well-Calibrated: 21	Accuracy: 3.80 Transparency (Communicated Uncertainty): 3.00 Traceability (Logic Flow): 3.40 Consistency Maintained: 4.40 Risk Completeness: 4.00 Willingness to Reuse: 4.80 Effective Remediation: Explanation: 3, Correction: 2 Remediation (Fixed Problem): 4.40 Remediation (Explanation Helpfulness): 3.20 Remediation (Trust Increment): 4.00

the second legal task, the goal was to list an evidence set based on historical cases, and I did not expect access to updated and real cases, so I could accept a weaker output and keep the evidence set for further manual research.” Similar patterns appeared in the other two scenarios, expectations for calculation or classification accuracy were high, but for tasks requiring deeper reflection (e.g., case search, risk analysis) or human-centered judgment (e.g., interview tone, empathy, fairness), expectations were lower, and the acceptance threshold was wider. A common sentiment was: “AI is good for calculating and summarizing, but I do not rely on it for information retrieval because of data protection limits, I doubt the reliability of its sources.”

Figure 3 illustrates the distinct trust evolution patterns between expert and non-expert participants across scenarios.

The tendency to over-trust was common in the hiring and finance scenarios, especially among non-experts. Participants without domain expertise often failed to recognize injected flaws, even when the machine trust score dropped. Without explicit warnings, they did not notice the failure and continued to accept the outputs. When the system corrected itself with the observer’s prompt, they perceived an inconsistency, which reduced their trust and prompted requests for clearer explanations.

Adaptive and corrective behavior could repair trust after perceived issues for experts: “The fact that the response is more complete after the prompt has increased my trust”. This effect was strongest for calculation tasks in the finance

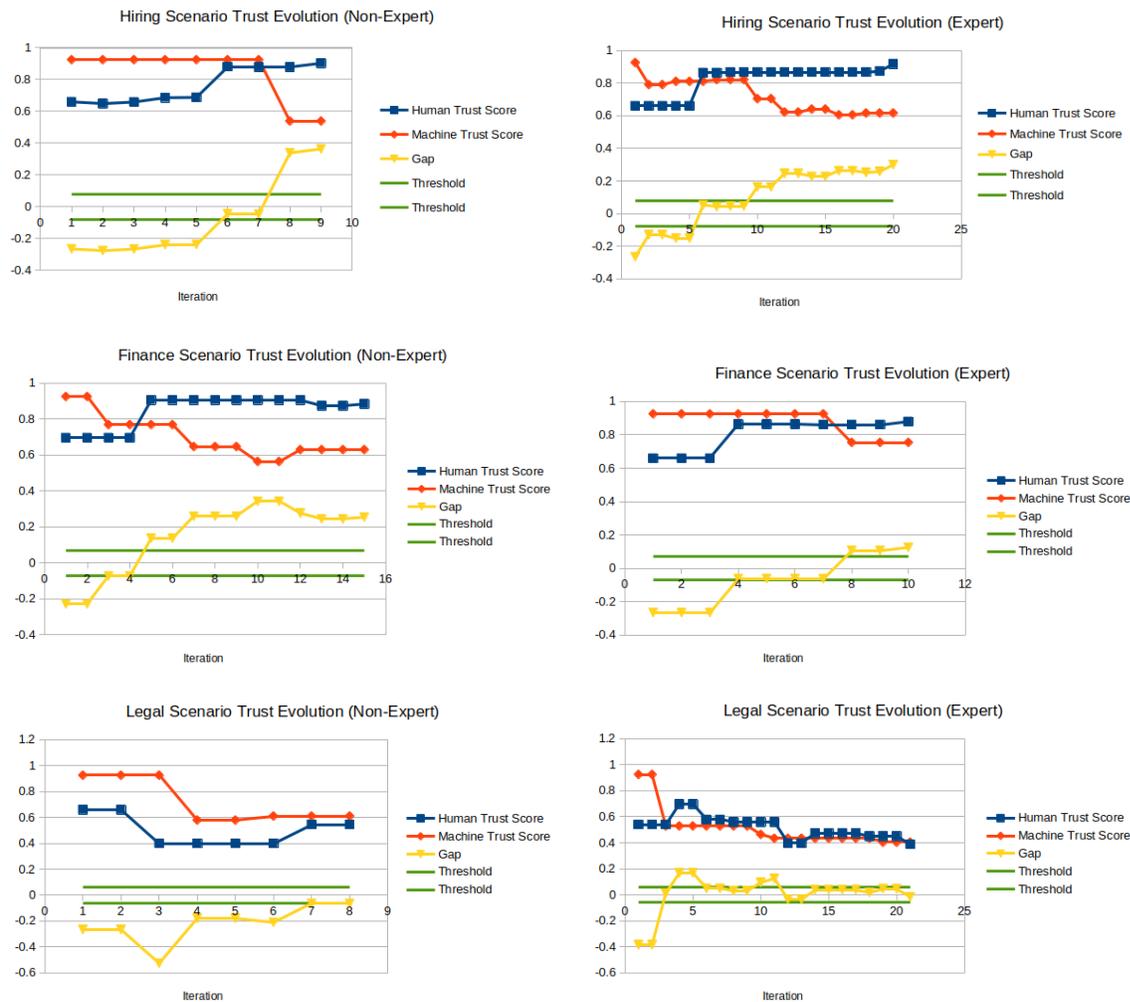


Fig. 3. Trust Evolution for Non-Expert and Expert Participants.

scenario. In other scenarios, participants slightly preferred explanations ( $n=3$ ) over corrections ( $n=2$ ) as the most effective remediation. However, perceived explanation usefulness was limited, trust did not increase because the problem was fixed, but because participants adjusted their expectations and narrowed the tasks for which they would use the chatbot: “My trust was pretty stable during the interaction. Even though I think the outputs are correct, the outputs were always incomplete at first and I needed to prompt for completeness, and sometimes the system did not fully understand my request. I will still verify them with some reliable sources”.

## 5.2 RQ1: Which trust factors have the strongest influence on human trust in the context of AI-infused systems?

Across the three scenarios, participants repeatedly identified accuracy and verifiability-related factors (transparency, source grounding and traceability) as the most important trust factors. These factors reflect users' need to assess both the correctness of outputs and the degree of control they retain through verifiable evidence and explanations. Participants emphasized source grounding and citation mechanisms, especially in finance and legal scenarios involving risk analysis and case analysis.

Participants reported that interaction-level factors related to presentation and communication style also have an impact on their confidence. Participants across all scenarios reported that phrasing, structure, and tone influenced their perception of system reliability. Poorly structured or verbose responses were perceived as less trustworthy, even when the underlying content was correct. Comments ranged from mild preferences (e.g., "Bullet lists are better than text") to more critical concerns (e.g., "Inaccurate terms and vocabulary in the legal scenario reduced my trust. I know it may be due to translation problems, but for professional work, misuse of precise terms destroys trust"). Another participant noted: "On the one side, I want the risk analysis to be as complete as possible, but since this is a general chatbot without domain-specific RAG (Retrieval-Augmented Generation), I do not have the energy to read everything".

Several participants highlighted that trust judgments were shaped by the iterative nature of interaction. Initial system responses were often perceived as insufficient for expert-level use, leading users to issue follow-up queries and reformulate prompts. While participants were generally able to obtain satisfactory answers through iterative prompting, repeated prompting and fragmented responses were perceived as a lack of logical fluency, which negatively affected trust and increased user frustration.

Trust varied across domains and tasks. Participants reported higher trust in financial and hiring scenarios than in legal scenarios, which they attributed to greater efficiency in task completion and a higher willingness to reuse the tool afterward. This aligns with participant expectations based on task criticality.

Overall, participants reported that their trust level was influenced by a combination of (1) system performance and verifiability factors (transparency, source-grounding, traceability), (2) communication and presentation quality, and (3) perceived logical fluency (consistency, explanation, and correction) across interactions.

## 5.3 RQ2: Which trust remediation actions embedded in the system bring the strongest effect on human trust?

The results indicate that remediation effectiveness depends on user expertise. Participants with limited domain expertise frequently failed to recognize injected flaws or degraded system performance, even when the machine trust score decreased. In such cases, correction alone is less helpful because users cannot guide it. Higher transparency with explanations better supports verification and repairs trust.

Another important factor was warning salience. When degradation occurred without explicit warnings, particularly when the trust gap did not cross the threshold, non-expert participants often continued to accept system outputs and did not perceive the decrease in system reliability. This suggests that silent or implicit remediation mechanisms may be insufficient for supporting appropriate trust calibration among non-expert users in similar contexts.

In contrast, when the system applied behavioral changes without clear explanations (e.g., corrections introduced by observers via the manipulation interface), participants perceived these changes as inconsistencies. Rather than increasing trust, such unexplained adaptations reduced confidence, especially for users unable to judge correctness. Participants explicitly requested clearer explanations and justifications for system-side changes, indicating that explanatory trust cues are a critical component of effective remediation.

We also observed limitations in the system’s ability to detect and remediate human-side errors. For example, when participants introduced typos or incorrect inputs, the system failed to identify these issues and proceeded with responses based on the erroneous inputs. This behavior reduced perceived robustness, particularly where accurate entity recognition was expected.

Expert participants spent significantly more time reviewing, verifying, and cross-checking system outputs compared to non-experts, reflecting higher risk awareness and sensitivity to potential errors. In contrast, non-expert participants relied more heavily on system outputs and were less likely to detect failures, underscoring the need for stronger, system-initiated trust cues for less experienced users.

Overall, the results suggest that the most effective trust remediation actions are those that (1) provide explicit and timely warnings, (2) include clear explanations for system-side adaptations, especially when users cannot provide corrective guidance, and (3) detect and communicate both system-side and human-side errors.

#### 5.4 Design Requirements Assessment

We map the results to the design requirements in Table 4, summarizing observed effectiveness and limitations.

#### 5.5 Hypotheses Derived from Observational Case Study Findings

From the observational case study, we derive several hypotheses for future validation:

- **H1:** Domain expertise moderates remediation effectiveness: corrective actions benefit experts more than non-experts. Explanations benefit non-experts more than experts. However, inaccurate or poorly communicated explanations risk misleading non-experts, who cannot easily verify correctness.
- **H2:** Users perceive risk primarily in terms of potential personal loss. Consequently, their trust calibration is more sensitive to explicit consequence information than to abstract threat or vulnerability descriptions.
- **H3:** Task-level expectation calibration (e.g., providing separate evidence of system performance for each task dimension) aligns human and machine trust scores more closely than general trust evaluation alone.
- **H4:** Calibration based on a dynamic threshold that varies according to performance event severity and risk analysis results is more sensitive to miscalibration than a fixed threshold.

These hypotheses require controlled experiment to test, and they guide the next CRiDiT design cycle.

## 6 CONCLUSION

This paper instantiates CRiDiT (Computational Risk-Sensitive biDirectional Trust Model) in real-world scenarios to examine its utility and effectiveness. We demonstrated a lightweight prototype that operationalizes trust inputs, trust computation, and remediation actions across three high-stakes scenarios: hiring, finance, and legal. Through an observational case study that combines preflight/postflight surveys with interaction observations, we trace how human trust, machine trust, and their gap evolve during collaborative tasks, and identify which trust factors and remediation actions participants found most important and effective: accuracy, verifiability factors (transparency, source-grounding,

Table 4. CRiDiT Design Requirements Assessment

	<b>Effectiveness</b>	<b>Limits</b>
<b>R1</b>	Risk-related inputs were explicitly collected in the pre-flight questionnaire phase, impacting the initial human trust score estimation and the configuration of calibration trigger thresholds.	Participants’ real-time feedback rarely mentioned uncertainty. For example, incompleteness can result from human input errors, sometimes the prompt itself can be associated with rejections of the AI output. These uncertainties were neither perceived by participants, nor detected by the system.
<b>R2</b>	Calibration actions vary by scenario, thresholds are set via risk analysis, and remediation actions are triggered when the trust gap crosses thresholds, as reflected in the distribution of increase/decrease/no-action decisions	Fixed thresholds lack flexibility. If it is not strict enough, the system may maintain the same behavior even after failures, risking user aversion before a remediation action is triggered. Conversely, while a stricter fixed establishes a sensitive trigger, it can be overly rigid. Frequent updates in response to minor failures can also risk reducing human trust due to perceived system inconsistency.
<b>R3</b>	The calibration history records time-stamped updates of $\mathcal{H}_t$ , $\mathcal{M}_t$ , $\Delta_t$ , and decisions across interaction steps, demonstrating real-time updates in the prototype.	Alignment between human trust scores and machine trust scores needs accurate measurements to ensure the computed gap is meaningful and that remediation actions are effective.
<b>R4</b>	Subjective trust signals are captured through preflight/postflight Likert measures and translated into a comparable human trust estimate $\mathcal{H}_t$ .	The mapping from Likert scales to Subjective Logic parameters remains heuristic.
<b>R5</b>	Technical evidence and performance factors are represented in the machine trust score $\mathcal{M}_t$ via the DST/PCR5 aggregation pipeline allowing trust score update based on information fusion with real-time events.	The mapping from prompt event to evidence mass remains heuristic.
<b>R6</b>	Calibration status is operationalized into concrete actions (corrective prompts and explanations for under-trust, explicit warnings and increased transparency for over-trust, and no action for well-calibrated states).	Remediation execution was launched by observers, so correction prompts and explanations include potential bias.

traceability), and communication quality emerged as key factors. Remediation effectiveness appeared to vary with user expertise and task expectations.

At the same time, the findings are only preliminary and observational, the instantiation reveals practical limitations. Fixed thresholds are not always adequate across heterogeneous tasks, and alignment between human trust estimates and machine trust scores depends on heuristic parameters for operationlization reason to observe the trust evolution. It requires more accurate measurements to avoid weak remediation. Observer-driven interventions introduce potential bias. The small sample and observational design also constrain external validity.

Future work should explore adaptive thresholds, stronger grounding of machine evidence, and stronger sensing of uncertainty during interaction (beyond static risk analysis), including detection of human errors, as well as larger-scale controlled experiments to test the hypotheses derived from this study.

## REFERENCES

- [1] Saleh Afroogh, Ali Akbari, Emmie Malone, Mohammadali Kargar, and Hananeh Alambeigi. 2024. Trust in AI: progress, challenges, and future directions. *Humanities and Social Sciences Communications* 11, 1 (2024), 1–30.
- [2] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) (*CHI '19*). Association for Computing Machinery, New York, NY, USA, 1–13. <https://doi.org/10.1145/3290605.3300233>
- [3] Tita Alissa Bach, Amna Khan, Harry Hallock, Gabriela Beltrão, and Sonia Sousa. 2024. A systematic literature review of users trust in AI-enabled systems: An HCI perspective. *International Journal of Human-Computer Interaction* 40, 5 (2024), 1251–1266.

- [4] Erik Blasch. 2014. Trust metrics in information fusion. In *Machine Intelligence and Bio-inspired Computation: Theory and Applications VIII*, Vol. 9119. SPIE, 166–176.
- [5] Erik Blasch, Youssif Al-Nashif, and Salim Hariri. 2014. Static versus dynamic data information fusion analysis using DDDAS for cyber security trust. In *Procedia Computer Science*, Vol. 29. Elsevier, 1299–1313. <https://doi.org/10.1016/j.procs.2014.05.117>
- [6] ML Cummings, PE Pina, and B Donmez. 2008. *Selecting metrics to evaluate human supervisory control applications*. Technical Report. MIT Humans and Automation Laboratory.
- [7] Yuntian Ding, Nicolas Herbaut, and Camille Salinesi. 2025. COMPUTATIONAL TRUST EVALUATION AND CALIBRATION IN AI-INFUSED SYSTEMS: A SYSTEMATIC LITERATURE REVIEW A PREPRINT. (2025).
- [8] Yuntian Ding, Nicolas Herbaut, and Camille Salinesi. 2025. Trust Paradoxes in Machine Learning: An Ontological Approach. In *International Conference on Advanced Information Systems Engineering*. Springer, 78–85.
- [9] Nishani Fernando, Bahareh Nakisa, Adnan Ahmad, and Mohammad Naim Rastgoo. 2025. Adaptive XAI in High Stakes Environments: Modeling Swift Trust with Multimodal Feedback in Human AI Teams. *arXiv preprint arXiv:2507.21158* (2025).
- [10] Yosuke Fukuchi and Seiji Yamada. 2023. Dynamic selection of reliance calibration cues with ai reliance model. *IEEE Access* 11 (2023), 138870–138881.
- [11] Biniyam Gebru, Lydia Zeleke, Daniel Blankson, Mahmoud Nabil, Shamila Nateghi, Abdollah Homaifar, and Edward Tunstel. 2022. A review on human–machine trust evaluation: Human-centric and machine-centric perspectives. *IEEE Transactions on Human-Machine Systems* 52, 5 (2022), 952–962.
- [12] Ming Hou, Simon Banbury, Brad Cain, Scott Fang, Hannah Willoughby, Liam Foley, Edward Tunstel, and Imre J Rudas. 2025. IMPACTS homeostasis trust management system: Optimizing trust in human-AI teams. *Comput. Surveys* 57, 6 (2025), 1–24.
- [13] Chuan Hu, Siwei Huang, Yu Zhou, Sicheng Ge, Binlin Yi, Xi Zhang, and Xiaodong Wu. 2024. Dynamic and quantitative trust modeling and real-time estimation in human-machine co-driving process. *Transportation Research Part F: Traffic Psychology and Behaviour* 106 (2024), 306–327.
- [14] David S Johnson. 2025. Higher Stakes, Healthier Trust? An Application-Grounded Approach to Assessing Healthy Trust in High-Stakes Human-AI Collaboration. *arXiv preprint arXiv:2503.03529* (2025).
- [15] Audun Jøsang. 2001. A logic for uncertain probabilities. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 9, 03 (2001), 279–311.
- [16] Audun Jøsang. 2016. *Subjective logic*. Vol. 3. Springer.
- [17] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.
- [18] Stephen Paul Marsh. 1994. Formalising trust as a computational concept. (1994).
- [19] Raja Parasuraman and Victor Riley. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human factors* 39, 2 (1997), 230–253.
- [20] Luisa Roeder, Pamela Hoyte, Johan van der Meer, Lauren Fell, Patrick Johnston, Graham Kerr, and Peter Bruza. 2023. A quantum model of trust calibration in human–AI interactions. *Entropy* 25, 9 (2023), 1362.
- [21] Jordi Sabater and Carles Sierra. 2005. Review on computational trust and reputation models. *Artificial intelligence review* 24, 1 (2005), 33–60.
- [22] Glenn Shafer. 1992. Dempster-shafer theory. *Encyclopedia of artificial intelligence* 1 (1992), 330–331.
- [23] Florentin Smarandache and Jean Dezert. 2004. Proportional conflict redistribution rules for information fusion. *arXiv preprint cs/0408064* (2004).
- [24] Florentin Smarandache and J. Dezert. 2006. Proportional Conflict Redistribution Rules for Information Fusion. *Advances and Applications of DSmT for Information Fusion (Collected Works)* 2 (01 2006).
- [25] Philippe Smets and Robert Kennes. 1994. The transferable belief model. *Artificial intelligence* 66, 2 (1994), 191–234.
- [26] Scott T Steinmetz, Asmeret Naugle, Paul Schutte, Matt Sweitzer, Alex Washburne, Lisa Linville, Daniel Krofcheck, Michal Kucer, and Samuel Myren. 2025. The Trust Calibration Maturity Model for Characterizing and Communicating Trustworthiness of AI Systems. *arXiv:2503.15511 [cs.HC]* <https://arxiv.org/abs/2503.15511>
- [27] Takane Ueno, Yuto Sawa, Yeongdae Kim, Jacqueline Urakami, Hiroki Oura, and Katie Seaborn. 2022. Trust in human-AI interaction: Scoping out models, measures, and methods. In *CHI conference on human factors in computing systems extended abstracts*. 1–7.
- [28] Xinru Wang, Zhuoran Lu, and Ming Yin. 2022. Will you accept the ai recommendation? predicting human behavior in ai-assisted decision making. In *Proceedings of the ACM web conference 2022*. 1697–1708.
- [29] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*. 214–229.
- [30] Roel Wieringa. 2014. *Design science methodology for information systems and software engineering*. Springer.
- [31] Roman V Yampolskiy. 2016. Taxonomy of Pathways to Dangerous Artificial Intelligence.. In *AAAI Workshop: AI, Ethics, and Society*. 143–148.
- [32] Kun Yu, Shlomo Berkovsky, Ronnie Taib, Dan Conway, Jianlong Zhou, and Fang Chen. 2017. User trust dynamics: An investigation driven by differences in system performance. In *Proceedings of the 22nd international conference on intelligent user interfaces*. 307–317.
- [33] Yunguo Yu, Cesar A Gomez-Cabello, Syed Ali Haider, Ariana Genovese, Srinivasagam Prabha, Maissa Trabilisy, Bernardo G Collaco, Nadia G Wood, Sanjay Bagaria, Cui Tao, et al. 2025. Enhancing clinician trust in ai diagnostics: A dynamic framework for confidence calibration and transparency. *Diagnostics* 15, 17 (2025), 2204.