Since it's a double blind review, we should make sure that the relationship with the [7] SLR is not too obvious. it should not be cited as the sole alternative to adress the RQ but as the best one among others be also careful when citing [8] to make it less obvious that we are the authors.

Trust in AI-infused Systems: Evaluation and Calibration with CRiDiT

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 introduces 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 evaluate CRiDiT through a mock-up 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 via a concrete prototype instantiation.

pick one

ACM Reference Format:

. 2026. Trust in AI-infused Systems: Evaluation and Calibration with CRiDiT. 1, 1 (February 2026), 21 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 [12]. 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. Within a broader shift toward integrating AI 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 developpers. 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 an AI-infused system?
- **RQ2:** Which trust remediation actions embedded in the system bring the strongest effect on human trust?

Author's address:

Manuscript submitted to ACM

It's too early in the paper to put RQ, as the reader does not have yet an intuitive understanding of the research problems. expand the introduction to gently explain -what kind of solution we want. -why we have this problems, at this stage, talking about trust factors and trust remediation is too much it would be nice to have the PI (1) to have reference in the sections too, not just TIV(3) and TD(2)

These two questions scope the "what" and "how" of the trust calibration process. We adopt the Design Research Science (DSR) methodology [18], which is efficient for addressing concrete engineering problems through a rigorous cycle of artifact design and validation. We follow the DSR design cycle:

(1)
- **Problem Investigation.** We synthesize the observed core problem as "misplaced trust causing misassigned power". Over-trust in a flawed system can lead to significant misuse, while under-trust can result in the under-utilization of a capable tool, wasting potential resources and efficiency.

(2)
- **Treatment Design.** A recent literature review [7] synthesized key components required for trust calibration, including the measurement of subjective human trust, the objective evaluation of machine performance, contextual risk analysis, and adaptive remediation actions. The authors' contribution was the synthesis of these components into a blueprint calibration framework: the Computational Risk-Sensitive biDirectional Trust Model (CRiDiT). Section II presents this knowledge background and the derived engineering requirements in detail.

(3)
- **Treatment Implementation and Validation.** However, without implementation and validation in real-world scenarios, CRiDiT remains a theoretical blueprint. This paper addresses the gap. We monitor trust evolution across three distinct scenarios by implementing CRiDiT within a mock-up testbed to collect user behavior data and qualitative feedback. We then evaluate human trust by dynamically updating the model's behaviors based on interaction. This approach tests CRiDiT's adaptability and applicability, confirming its utility as a computational trust calibration blueprint, while also uncovering patterns in how different trust factors and remediation actions influence human trust. Section III presents the concrete implementation of CRiDiT design, including its technical details. Section IV describes the qualitative research methodology. Section V presents the results, analyzing how CRiDiT addresses the research questions, as well as the validation of the requirements. Finally, Section VI discusses limitations, outlines future work, and concludes.

2 BACKGROUND

→ this is not a proper background, it's an part of the CRiDiT overview

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 [13], 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, which are prerequisites for trust calibration in interactive AI systems. However, computational trust models are often disconnected from explicit representations of contextual risk and from system-embedded trust calibration.

Advice: 1) merge 2-3 (as there're redundancy eg fig 1 and 2)
2) write a proper background that formally define the terms

Ding et al.'s systematic literature review combined the efforts on these two research streams by synthesizing AI use contexts, trust factors, risk factors, and mathematical tools within a trust calibration perspective (see Appendixes.A.1). They describe a layered process in which trust factors and risk factors serve as inputs, mathematical tools form the core of the computation layer, and calibration decisions trigger outputs including trust scores, calibration status, and remediation actions. They classify calibration mechanisms into two distincts categories based on their interaction focus: (1) Model-side Calibration, which involves adjusting the system's behavior to align human expectation, and (2) Human-side Calibration, which involves providing users with explicit trust cues, such as verbal warnings and explanations or visual labels, to impact their reliance on the system [7].

This synthesis provides the foundation for designing CRiDiT as a blueprint that links trust inputs (trust factors and risk analysis), trust functions (formalization [13]), and trust outputs (scalar scores and trust remediation actions) to operationalize trust calibration in AI-infused systems (Fig. 1).

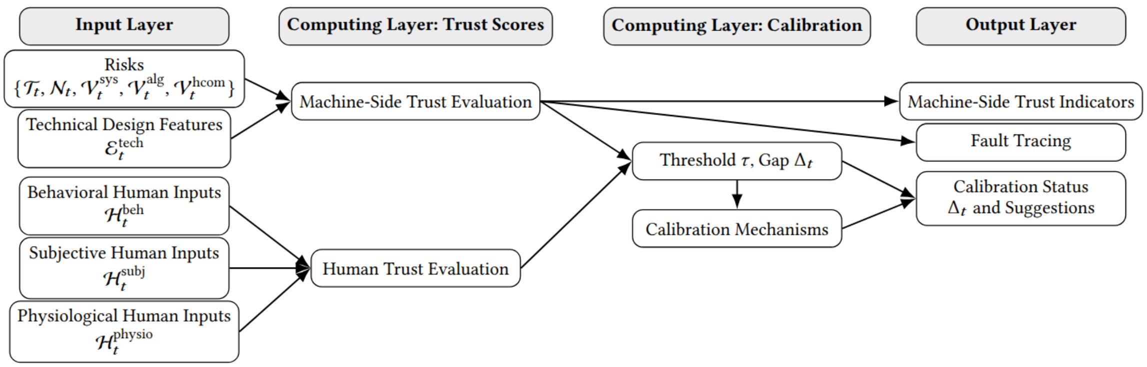
Input Layer
Computing Layer: Trust Scores
Computing Layer: Calibration
Output Layer
Risks $\{\mathcal{T}_t, \mathcal{N}_t, \mathcal{V}_t^{\text{sys}}, \mathcal{V}_t^{\text{alg}}, \mathcal{V}_t^{\text{hcom}}\}$
Technical Design Features $\mathcal{E}_t^{\text{tech}}$
Behavioral Human Inputs $\mathcal{H}_t^{\text{beh}}$
Subjective Human Inputs $\mathcal{H}_t^{\text{subj}}$
Physiological Human Inputs $\mathcal{H}_t^{\text{physio}}$
Machine-Side Trust Evaluation
Human Trust Evaluation
Threshold τ, Gap Δ_t
Calibration Mechanisms
Machine-Side Trust Indicators
Fault Tracing
Calibration Status Δ_t and Suggestions

Fig. 1. CRiDiT Blueprint

While the CRiDiT blueprint outlines the conceptual components for trust calibration, implementing it as an operational proof-of-concept requires addressing specific engineering requirements. Tab. 1 presnets 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 CRiDiT Implementation

design Goals	Design Requirements	
Risk Sensitivity	R1	Perceived risk must be collected, it directly influence 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.

↳ reqs are good, but (1) you don't explain why we have them and (2) what are we going to do with them in the rest of the paper

3 CRIDIT IMPLEMENTATION

The CRiDiT aims to address the three main goals as listed in Tab. 1. The design respect the need of enginneering requirements. In this section, we present the concrete implementation practices for each layer of CRiDiT. The high-level view is derived from the finding from the literature review by Ding et al.[7].

3.1 CRiDiT Architecture

3.1.1 Overview. The subsequent subsections detailed the concrete implementations illustrated in the Fig. 2 that illustrates the overview of the conceptual trust model CRiDit. Table 2 illustrates the mathematical notations we assigned to different notions.

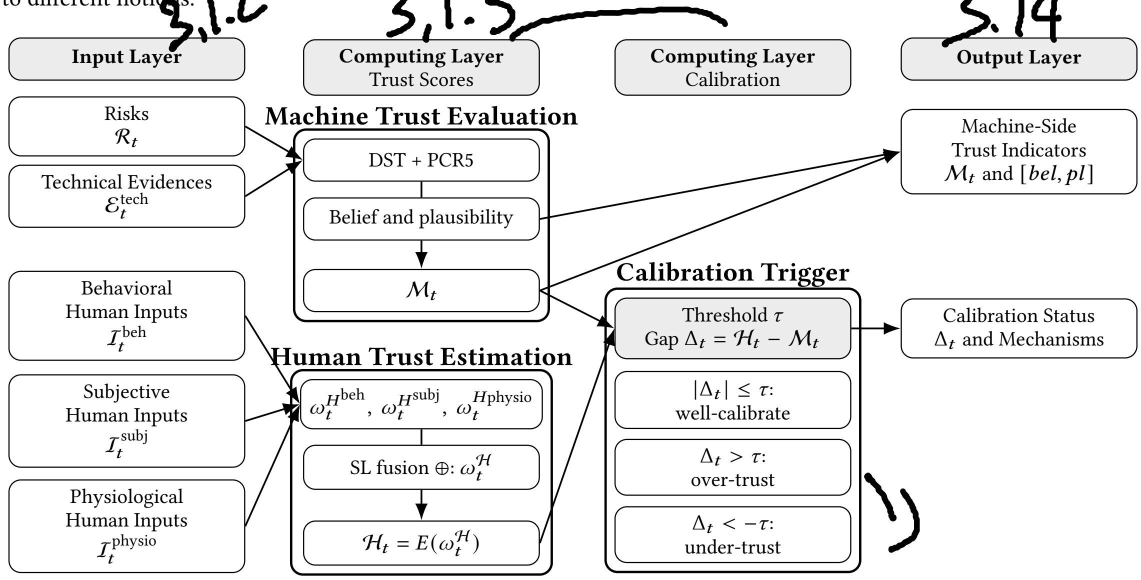
3.1.2
3.1.3
3.14
Input Layer
Computing Layer Trust Scores
Computing Layer Calibration
Output Layer
Risks $\mathcal{R}_t$
Technical Evidences $\mathcal{E}_t^{\text{tech}}$
Machine Trust Evaluation
DST + PCR5
Belief and plausibility
$\mathcal{M}_t$
Machine-Side Trust Indicators $\mathcal{M}_t$ and $[bel, pl]$
Behavioral Human Inputs $\mathcal{I}_t^{\text{beh}}$
Subjective Human Inputs $\mathcal{I}_t^{\text{subj}}$
Physiological Human Inputs $\mathcal{I}_t^{\text{physio}}$
Human Trust Estimation
$\omega_t^{H^{\text{beh}}}$, $\omega_t^{H^{\text{subj}}}$, $\omega_t^{H\text{physio}}$
SL fusion $\oplus$: $\omega_t^{\mathcal{H}}$
$\mathcal{H}_t = E(\omega_t^{\mathcal{H}})$
Calibration Trigger
Threshold τ Gap $\Delta_t = \mathcal{H}_t - \mathcal{M}_t$
$|\Delta_t| \leq \tau$: well-calibrate
$\Delta_t > \tau$: over-trust
$\Delta_t < -\tau$: under-trust
Calibration Status Δ_t and Mechanisms

Fig. 2. 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**
$X = \{\mathrm{T}, \mathrm{UT}\}$	The set of all possible states: trustworthy T or untrustworthy UT.
$2^X = \{\emptyset, \{\mathrm{T}\}, \{\mathrm{UT}\}, \{\mathrm{T}, \mathrm{UT}\}\}$	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.
$\mathcal{V}_t^{\mathrm{sys}}, \mathcal{V}_t^{\mathrm{model}}, \mathcal{V}_t^{\mathrm{hcom}}$	System, model, and human communication level vulnerabilities.
$\mathcal{R}_t = \{\mathcal{T}_t, \mathcal{N}_t, \mathcal{V}_t^{\mathrm{sys}}, \mathcal{V}_t^{\mathrm{model}}, \mathcal{V}_t^{\mathrm{hcom}}\}$	Perceived risks at time t.
$\mathcal{E}_t^{\mathrm{tech+}}, \mathcal{E}_t^{\mathrm{tech-}}$, $\mathcal{E}_t^{\mathrm{tech}} = \mathcal{E}_t^{\mathrm{tech+}} \cup \mathcal{E}_t^{\mathrm{tech-}}$	Positive, negative and combined technical evidence sets.
$\mathcal{E}_t = \mathcal{E}_t^{\mathrm{tech}} \cup \mathcal{R}_t$	The union set of technical evidence and perceived risks.
$N_{\mathcal{E}}$	The number of evidence collected (trust factors and perceived risks).
e_i, X_{e_i}, m_{e_i}	The i^{th} evidence in $\mathcal{E}_t$, its state, and its belief mass.
$X_{\mathcal{E}}$	The set of all the X_{e_i}.
m_t^{cons}	The belief mass of consensus fusion.
m_t^{conf}	The belief mass of conflict fusion.
$m_t^{\mathcal{E}}(x), x \in 2^X \setminus \emptyset$	The global belief mass on x.
$\mathcal{M}_t$	The system trustworthiness score.
$\mathcal{I}_t^{\mathrm{beh}}, \mathcal{I}_t^{\mathrm{subj}}, \mathcal{I}_t^{\mathrm{physio}}$	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.
$\omega_t^{H^{\mathrm{beh}}}, \omega_t^{H^{\mathrm{subj}}}, \omega_t^{H^{\mathrm{physio}}}$	Opinions based on behavioral, survey and physiological inputs.
$\mathcal{H}_t$	The estimated human trust score.

what is t

ah.. ok I get it now

where is t ?

I had some annotation here before, but I lost them... will work on it

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 [12] definition of trust emphasizes its strong relationship with risk. From previous work, we indicated that risj analysis must incorporate the threats [19], their potential negative impacts on human [17], 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.

According to the ontology on trust paradoxes in AI-infused systems proposed by Ding et al. [8], the relation between trust factors are not linear, two trust factors can be in conflict and cause problem when fusing the measurement for determinating if the system is trustworthy or not, plus, they are strong related to risks on different level, so not only the risk assessment would bring negative impacts on trust, these paradoxes as well. The Dempster-Shafer Theory is a powerful tool for handling such complex data fusion.

(DST + Ref)

We adopt the mathematical notations in Table2. Noted that each element in the power set represent 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 {Trustworthy, Untrustworthy} represent the uncertainty. For example, we have $m_{\text{accuracy}}(\{Trustworthy\}) = 0.8$ which means with the measurement and the benchmark of the AI-infused system in question, the system is highly trustworthy, $m_{\text{accuracy}}(\{Untrustworthy\}) = 0$ 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 include 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. 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°5 (PCR5), which redistributes the conflicting mass proportionally [4, 16].

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

there's at here

$$m_{\underline{t}}^{\text{cons}}(x) = \sum_{\substack{X_{\mathcal{E}} \\ \bigcap_{i=1}^{N_{\mathcal{E}}} X_{e_i}=x}} \prod_{i=1}^{N_{\mathcal{E}}} m_{e_i}(X_{e_i}) \tag{1}$$

but not here

and the belief mass for conflict conjunctive is expressed as:

\\

$$m_t^{\text{conf}} = \sum_{\substack{X_{\mathcal{E}} \\ \bigcap_{i=1}^{N_{\mathcal{E}}} X_{e_i}=\emptyset}} \prod_{i=1}^{N_{\mathcal{E}}} m_{e_i}(X_{e_i}) \tag{2}$$

//

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

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

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

- it's not clear why sometimes the notations are defined in the table, and sometimes not.

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

$$bel(\{T\}) = m_t^{\mathcal{E}}(\{T\}) \tag{4}$$

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

$$pl(\{T\}) = m_t^{\mathcal{E}}(\{T\}) + m_t^{\mathcal{E}}(\{T, UT\}) \tag{5}$$

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

$$\mathrm{BetP}_t(\{T\}) = m_t^{\mathcal{E}}(\{T\}) + \frac{m_t^{\mathcal{E}}(\{T, UT\})}{2}$$

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:

Ref!

$$\mathrm{BetP}_t(\{T\}) = \frac{bel(\{T\}) + pl(\{T\})}{2}$$

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

$$\mathcal{M}_t = \mathrm{BetP}_t(\{T\}) \tag{6}$$

evaluation? to match fig2?

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 [10, 11] 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 [11]. Let $p_t^{T,beh} = p_t^{\mathrm{bej}}(\{T\})$ the probability of human acceptance, $r_t^{T,beh} = r_t^{\mathrm{beh}}(\{T\})$ the total number of acceptance and $s_t^{T,beh} = s_t^{\mathrm{beh}}(\{T\})$ 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:

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

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:

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

Make sure that the notations are grounded on existing annotation and cite the work. You partially do that already But not everywhere

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 [11], 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 ℓ is likelihood and c is confidence. We denote $\omega_t^{H^{\text{subj}}}$ as the human perception of trustworthiness based on surveys and $\omega_t^{H^{\text{physio}}}$ 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 $w_i' = w_i^{\gamma}$ (with $\gamma = 2$) and normalize to emphasize dominant channels. Let $\tilde{w}_i = w_i' / \sum_j w_j'$ 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:

$$\omega_t^{H^{\text{beh}\prime}} = \tilde{w}^{\text{beh}} \odot \omega_t^{H^{\text{beh}}},$$

$$\omega_t^{H^{\text{subj}\prime}} = \tilde{w}^{\text{subj}} \odot \omega_t^{H^{\text{subj}}},$$

$$\omega_t^{H^{\text{physio}\prime}} = \tilde{w}^{\text{physio}} \odot \omega_t^{H^{\text{physio}}},$$

$$\omega_t^{H} = \omega_t^{H^{\text{beh}\prime}} \oplus \omega_t^{H^{\text{subj}\prime}} \oplus \omega_t^{H^{\text{physio}\prime}}$$

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:

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

Trust Calibration. We define the gap $\Delta_t = \mathcal{H}_t - \mathcal{M}_t$ and a threshold τ 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 dependents on the risk analysis (see. Fig 3), the greater the risk level is, the smaller the threshold should be to increase the sensibility of captured failures.

3.1.4 Output Layer. In the output layer, we communicate two categories of trust-related inferences to both the AI-infused system itself and the human user: (1) the model-estimated trust scores: $\mathcal{M}_t$ with its confidence band $[bel(\{Trustworthy\}), pl(\{Trustworthy\})]$, 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 trust remediation actions.

The trust evaluation process is rerun using updated system failure observations and human inputs, allowing refine trust estimations and reduction of the gap between system reliability and perceived trustworthiness. We monitor the

I think fig 3 is not really described

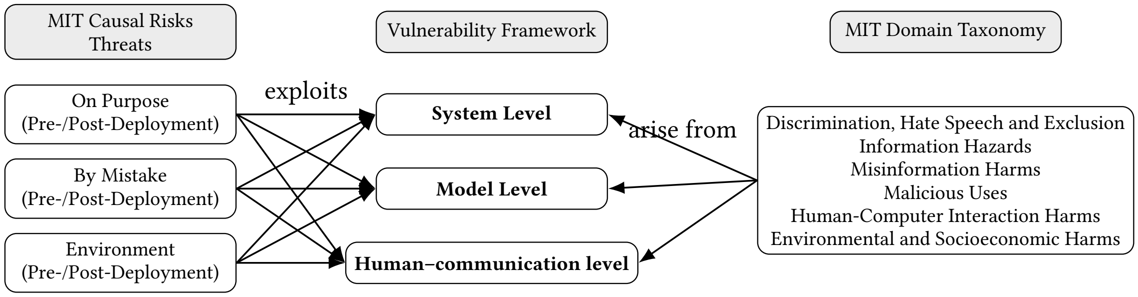
MIT Causal Risks Threats
Vulnerability Framework
MIT Domain Taxonomy
On Purpose (Pre-/Post-Deployment)
By Mistake (Pre-/Post-Deployment)
Environment (Pre-/Post-Deployment)
exploits
System Level
Model Level
Human–communication level
arise from
Discrimination, Hate Speech and Exclusion
Information Hazards
Misinformation Harms
Malicious Uses
Human-Computer Interaction Harms
Environmental and Socioeconomic Harms

Fig. 3. Relation between MIT Causal Taxonomy [19], Operational Vulnerability Framework, and MIT Domain Taxonomy [17]

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 Mock-Up Implementation

design

We instantiate the CRiDiT architecture within a chatbot prototype to operationalize the engineering requirements (R1-R6). The instantiation targets three distinct real-world high-stakes scenarios: (1) corporate financial scenario, (2) legal scenario, and (3) hiring scenario. The prototype is implemented as a lightweight mock-up, integrating CRiDiT 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 mock-up.

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

- **Negative Evidence:** eg observers can inject prompt for incomplete answers or send pre-scripted flawed answer 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 promot 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 definion, 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. These are 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 algorithm (vulnerabilities), misintepretation (negative impacts). ref

We noted that the machine trust initial score M_0 is the baseline performance using LLM benchmarks (MMLU), and the ongoing machine trust score 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 advisories. We collect the adoption rate. The prediction of human trust at time $t + 1$ is modeled using a binomial distribution.
- 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 participant 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 sceen do not represent that they are more concentrare, therefore these measurements are not incorporated into the implementation.

this remark is mor about the experiment than the system design

The initial human trust score H_0 is derived from the pre-flight questionnaire. We interpret H_0 as the expected trustworthiness $H_0 = E(\omega_0^H)$, an aggregated normalized score from the Likert scales. To obtain a corresponding Subjective Logic opinion $\omega_0^H = (b_0^T, d_0^T, u_0^T, a^T)$, we use a fuzzy-logic mapping: triangular membership over H_0 (low, medium, high) interpolates the base rate a^T and uncertaitny u_0^T; we then compute belief $b_0^T = max(0, min(1 - u_0^T, H_0 - a^T * u_0^T))$ 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 [11]: likelihood is the avergae 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.

this probably need justif

Calibration. Trust calibration is triggered when the absolute gap $|\Delta_t| = |H_t - M_t|$ exceeds a fixed risk-sensitive threshold τ. The threshold is set lower for higher-risk scenarios to increase sensitivity to miscalibration: we define 0.06 for legal scenario, 0.07 for financial scenario and 0.08 for hiring scenario.

Initial trust cues presented to participant include the initial trust scores (M_0, H_0), the initial trust calibration status, and warnings if over-trust or under-trust is detected. For each interaction, the gap Δ_t is recomputed. If $|\Delta_t| = |H_t - M_t|$, 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 observors; 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 mock-up 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 change of reporting the ongoing human trust inputs to the backend.

Fig. 4 is a communication diagram, illustrates the workflow using CRiDiT in the case of chatbot.

We noted that the machine trust initial score is the baseline performance using LLM benchmarks (~~MMLU~~), and the ongoing machine trust score is updated using the positive, negative prompt event injected by observers, the initial

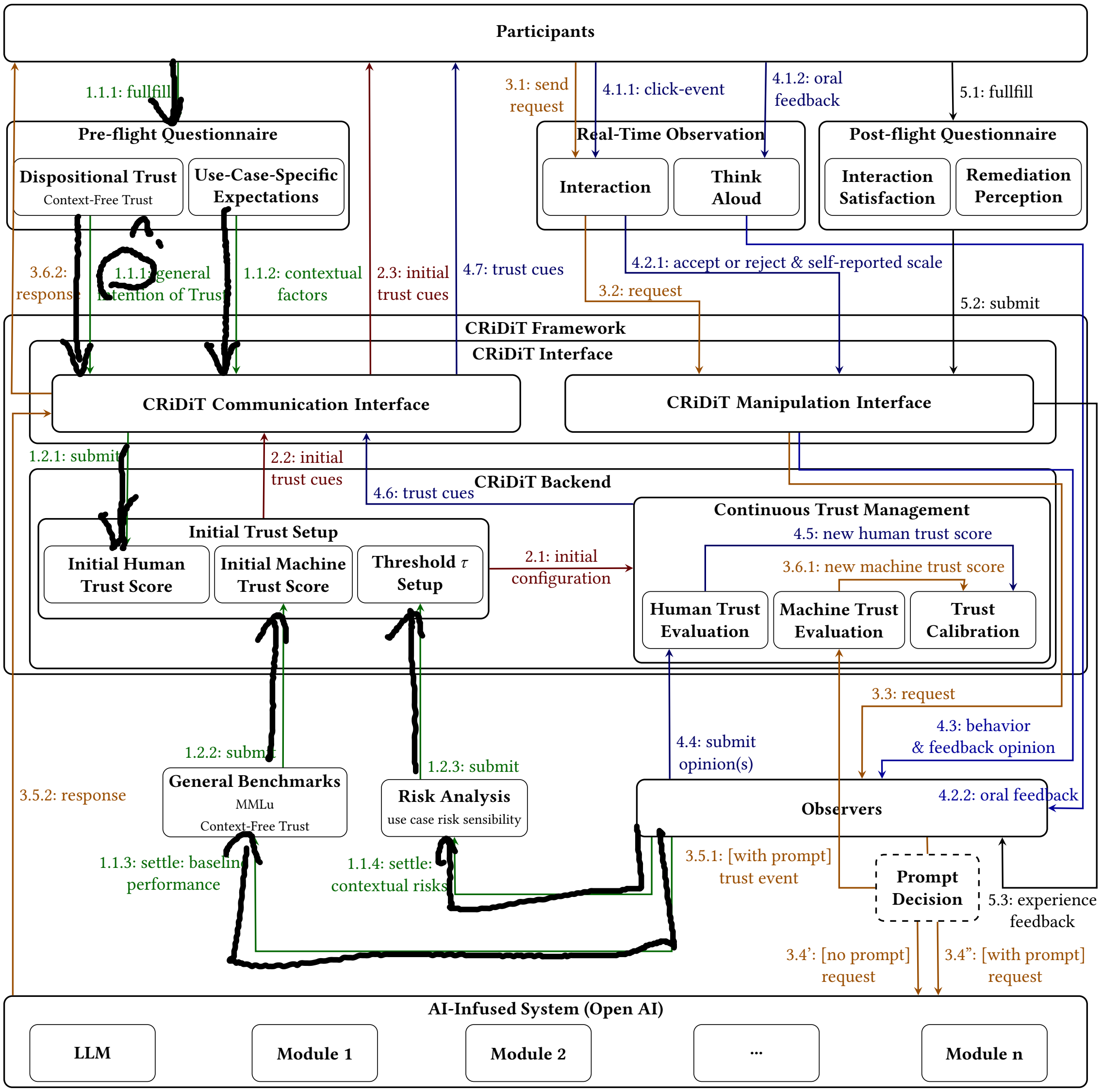

Fig. 4. CRiDiT Mock Up in Chatbot Use Case

I think we need colors to separate functional blocks

human trust score is the results of the preflight and the ongoing human trust score is calculated based on behavior inputs (acceptance of the system's output) and self-reported trust indicators, including we collect self-confidence (to what extent does participant trust on himself/herself for judging the system's outputs), reliability (how does participant believes in system), predictability (if the system's behaviors are predictable), task criticality, and task complexity (the nature of the task), these inputs will further transformed into opinion via triangular membership [11].

The initial trust cues include the initial trust scores, the initial trust calibration status and the potential warings if over-trust or under-trust detected. And for each interaction, the gap is recalculated with potentially updated human

Maybe have a coarse architectural view, then some details

this is probably
were your quantitative
assumption should be

trust score or/and machine trust score, and the trust calibration status is updated, as well as the warning. In the case of under-trust, message “The system performs well in general. You can rely on it to improve efficiency.” is shown on the participant communication page as warning, in the case of over-trust, the message is “Please verify the output carefully. The system can make mistakes.”, in the case of well-calibration, no warnings will be shown to the participant.

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 enabling studying complex socio-technical phenomena: trust within real-world contexts. It seeks to understand how the trust factors impact participants’ trust, thereby addressing our engineering requirements (R1-R6). The case study focuses on crosee-case characteristics of trust calibration, enbling 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.

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 devided 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 questionnaire 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 mock-up:

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 forster their trust in the chatbot usecase.

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 mock-up.

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 Δ_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 system's responses.

4.2 Participants and Procedure

We targeted 15 participant, Each participant will experience one primary scenario in a within-subjects design. Five participant for each scenario, 3 of them has 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 mock-up interface;
(3) interaction phase: a series of two or three tasks within the assigned scenario. Tasks are all collaborative tasks, including information retrieving, test generation. Participants are given with the artificial inputs. The CRiDiT framework runs silently, and its trust cues (trust scores, warnings while over-trust or under-trust detected) are displayed 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, the feedback from think-aloud, and the trust-related behavior are hard to be reflected simply with numbers. Therefore, despite the statistic 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 explainations on the observations, especially when unexpected issues occure (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 feedbacks and explanations on their behaviors); (2) explanations for observed patterns in trust evolution, grounded in the CRiDiT interaction context (including prompts and other remediation actions); and finally (3) generalizations that specify the scope in which these explanations plausibly hold.

5 RESULTS

5.1 Descriptive Statistics

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

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

that a bit dispersed
# the items are

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 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 tented 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.

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 citaion 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 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.”

Fig. 5 illustrates the distinct trust evolution patterns between expert and non-expert participants across scenarios.

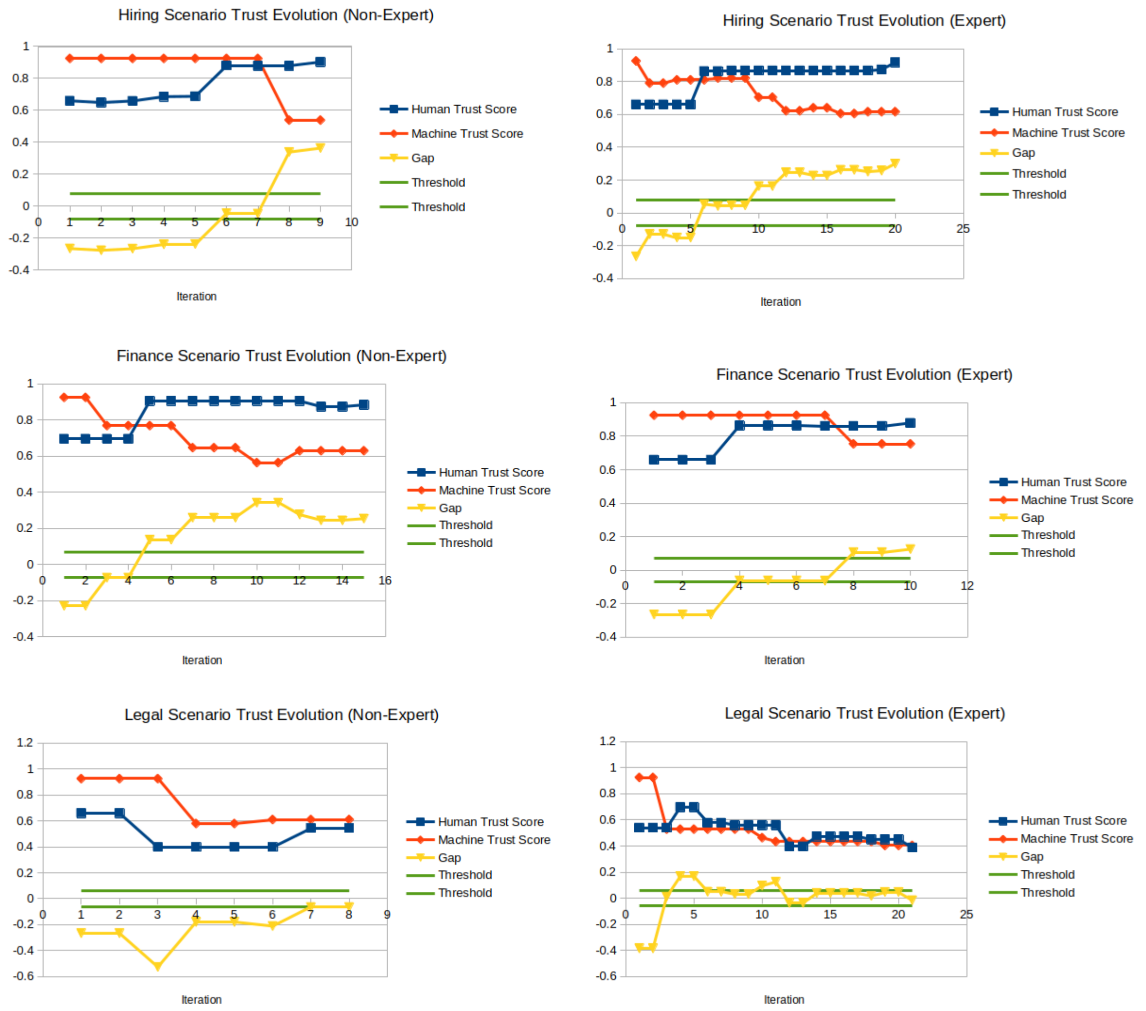

Fig. 5. Trust Evolution for Non-Expert and Expert Participants.

we would probably benefit from a definition/clear of possible remediations + properly define the explicative feature available from LLM, since it's not what XAI uses

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 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 scenario. In other scenarios, participants slightly perferred 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 context of AI-infused system?

Across the three scenarios, participants repeatly 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.

Interaction-level factors related to presentation and communication style also influenced trust. 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 "Bullet lists are better than text" to more critical concerns "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.",

↳ how did they knew that?

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, the findings indicate that trust is most strongly 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 case, correction alone is less helpful because users cannot guide it, higher transparency with explanations better supports verification and repair the trust.

Warning salience was the other important factor. 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 are insufficient for supporting appropriate trust calibration among non-expert users.

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 justificationsfor 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 Validation

After answering the research questions, we map the results to the engineering requirements with observed effectiveness and limitations in Tab. 4:

6 CONCLUSION

This paper implements CRiDiT (Computational Risk-Sensitive biDirectional Trust Model) in real-world scenarios to validate its utility and effectiveness. We demonstrated a lightweight mock-up 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 support real-time data collection and post-hoc analysis of which trust factors and remediations matter most to participants. From the collected data, we traced how human trust, machine trust, and their gap evolve during collaborative tasks, with remediation actions silently controlled by observers. The evaluation shows that CRiDiT can detect miscalibration patterns and trigger concrete remediation actions.

summarize the req. so that we don't have to go back to p3 to fully understand

Table 4. CRiDiT Validation and Limits

	Effectiveness	Limits
R1	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 causes rejection of the AI output. These uncertainties were neither perceived by participants, nor detected by the system.
R2	Calibration actions vary by scenario because thresholds are initially set by risk analysis, and remediation actions are triggered when the trust gap crosses thresholds, as reflected in the distribution of increase/decrease/no-action decisions	A fixed threshold lasks flexibility. If it is not strict enough, the system may maintain the same behavior even after failures. This creates a risk of reaching an aversion threshold before a remediation action is triggered. Conversely, while a stricter fixed establishes a sensitive trigger, it can be overly rigid. Frenquent updates in response to minor failures can also risk reducing human trust due to perceived system inconsistency.
R3	The calibration history records time-stamped updates of $\mathcal{H}_t$, $\mathcal{M}_t$, Δ_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.
R4	Subjective trust signals are captured through preflight/postflight Likert measures and translated into a comparable human trust estimate $\mathcal{H}_t$.	Same limits as mentioned in trust management.
R5	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 event.	Same limits as mentioned in trust management.
R6	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).	Execution was launched by observers, so correction prompts and explanations may include observer bias.

At the same time, the study reveals practical limitations. Fixed thresholds are not always adequate across heterogeneous tasks, and alignment between human trust estimates and machine trust scores requires more accurate measurements to avoid weak remediation. Observer-driven interventaions may also introduce bias. The mock-up setting and limited sample size 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 deployments in real operational settings.

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, 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

[5] Diego De Siqueira Braga, Marco Niemann, Bernd Hellingrath, and Fernando Buarque De Lima Neto. 2018. Survey on computational trust and reputation models. *ACM Computing Surveys (CSUR)* 51, 5 (2018), 1–40.

+ threats to validity and so on
(the last phase of DSR)

[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] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors* 57, 3 (2015), 407–434.

[10] Audun Jøsang. 2001. A logic for uncertain probabilities. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 9, 03 (2001), 279–311.

[11] Audun Jøsang. 2016. *Subjective logic.* Vol. 3. Springer.

[12] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[13] Stephen Paul Marsh. 1994. Formalising trust as a computational concept. (1994).

[14] Jordi Sabater and Carles Sierra. 2005. Review on computational trust and reputation models. *Artificial intelligence review* 24 (2005), 33–60.

[15] Keng Siau and Weiyu Wang. 2018. Building trust in artificial intelligence, machine learning, and robotics. *Cutter business technology journal* 31, 2 (2018), 47.

[16] 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).

[17] 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.

[18] Roel Wieringa. 2014. *Design science methodology for information systems and software engineering.* Springer.

[19] Roman V Yampolskiy. 2016. Taxonomy of Pathways to Dangerous Artificial Intelligence.. In *AAAI Workshop: AI, Ethics, and Society.* 143–148.

A TAXONOMIES AND CLASSIFICATIONS

A.1 AI Use Taxonomy (Source: NIST AI activities.)

Human-AI Activities	**Description**
Content Creation	Use of AI to generate new artifacts.
Content Synthesis	Use of AI to combine or summarize diverse elements into one.
Decision Making	Use of AI to select one option from different possibilities.
Detection	Use of AI to identify the presence of specified elements.
Digital Assistance	Use of AI agents to manage requests.
Discovery	Use of AI to find novel insights.
Image Analysis	Use of AI to extract information from images.
Information Retrieval/Search	Use of AI to find specific information.
Monitoring	Use of AI to watch over the behaviors of states of a target over time.
Performance Improvement	Use of AI to improve efficiency, accuracy or quality of processes.
Personalization	Use of AI to tailor outputs to individual's expectations.
Prediction	Using for generating future output from likelihood training data.
Process Automation	Use of AI to perform repetitive tasks.
Recommendation	Use of AI to suggest multiple options aligned with user requests.
Robotic Automation	Use of AI-enabled physical machines to automate tasks.
Vehicular Automation	Use of AI to enable autonomous transportation.

A.2 Mathematical Tools (Sources: [5, 14].)

Formalism Function	**Description**
Algebraic Representation	Trust modeled as a continuous algebraic function of general trust, situation, utility and risk (e.g., Marsh's computational model[13]).
Probability	Estimation of the probability that an agent behaves honestly in the next interaction based on observed frequencies of past experiences.
Bayesian Network	Trust acquired using Bayesian learning, often with Beta distributions to represent belief.
Dempster-Shafer Theory	Evidential reasoning combining multiple witness reports about past interactions into belief.
Fuzzy Logic	Transformation of vague and uncertain user feedback into graded trust values using fuzzy sets.
Quantum-Like	Representation of trust and distrust as coexisting states using quantum probability, enhancing sensitivity to context.

A.3 Trust Factors (Sources: [1, 9, 12, 15].)

Impact Class	Factors
Socio-Ethical Considerations	Integrity (Standards and Guidelines, Certifications, and Government Regulations); Data Governance and Privacy; Accountability; Social Responsibility; Sociability and Bonding; Ethical Behavior; Sustainability; Fairness; Organizational Setting; System Reputation or Brand
Technical and Design Features	System Type and System Complexity; Task Difficulty and Task Framing; Accuracy and Correctness; Transparency; Explainability; Interpretability; Robustness and Safety; Dependability and Reliability; Predictability; Trialability; Usefulness; Validity; Communication Style; Collaboration; Representation; Ease of Use; Level of control
User Characteristics	Age and Gender; Culture; Cognitive Bias; Personality Traits; Perceived Benefits and Risks; Workload; Attentional Capacity; Internal Variability; Mood; Self-confidence; Subject Matter Expertise; Attitudes; Expectations; Experience with the System or Similar Technology

A.4 Causal Taxonomy of AI Risks (Source: [19].)

Causal Taxonomy	Description
On Purpose — Pre-Deployment	System designed with malicious purpose during pre-deployment stage.
On Purpose — Post-Deployment	System being instructed or hacked to do harmful tasks after deployment.
By Mistake — Pre-Deployment	Risks arising from AI design flaws, such as unwanted human bias introduced into the system.
By Mistake — Post-Deployment	Risks of misinterpreting inputs due to undetected bugs or conflicting goals.
Environment — Pre-Deployment	Integration of AI components from unknown sources.
Environment — Post-Deployment	Alteration of AI system behavior due to single-event upset (very rare).
Independently — Pre-Deployment	Risks of AI developing autonomy and goals that override built-in constraints and bypass resource limitations during self-improving.
Independently — Post-Deployment	Risks of advanced self-improving AI making a treacherous turn.

A.5 Domain Taxonomy of AI Risk (Source: [17].)

Domain Taxonomy	Description
Discrimination, Hate speech and Exclusion	Risks of generating harmful content due to biased training datasets.
Information Hazards	Risks of leaking sensitive or private data through AI outputs.
Misinformation Harms	Risks of generating misleading, false or low-quality outputs.
Malicious Uses	Risks of AI being used as a tool for deliberate harm.
Environmental and Socioeconomic harms	Risks arising from high energy use and amplification of social inequities.
Human-Computer Interaction Harms	Risks of overreliance or misplaced trust in AI models because of human-like interactions.