

A Quantum-Inspired Framework for Real-Time Trust Estimation in Human-AI Interaction: A Pilot Study

Reyansh Badhwar

University of Calgary

Zehaan Walji

University of Calgary

Parshva Dave

University of Calgary

Junho Park

`junho.park@ucalgary.ca`

University of Calgary

Article

Keywords: Trust, Quantum cognition, Human-AI interaction, Conversational AI, Trust dynamics

Posted Date: March 9th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-8464028/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Scientific Reports on August 29th, 2026.
See the published version at <https://doi.org/10.1038/s41598-026-67931-x>.

A Quantum-Inspired Framework for Real-Time Trust Estimation in Human-AI Interaction: A Pilot Study

Reyansh Badhwar^a, Zehaan Walji^b, Parshva Dave^b, Junho Park^{c*}

a. Department of Software Engineering, Schulich School of Engineering, University of Calgary, Calgary, AB, Canada

b. Department of Computer Science and Biological Science, Faculty of Science, University of Calgary, AB, Canada

c. Department of Biomedical Engineering, Schulich School of Engineering, University of Calgary, Calgary, AB, Canada

*Corresponding author:

Junho Park, PhD

Office 212, Calgary Centre for Innovative Technology (CCIT) building

University of Calgary

2500 University Dr NW

Calgary, AB, T2N 4V8, Canada

Office: +1-403-220-5358

Email: junho.park@ucalgary.ca

Fax: N/A

Manuscript Category: Research Article

Abstract

Trust is a determinant of effective human-AI collaboration, shaping whether users appropriately accept or reject system recommendations. In safety-critical domains, miscalibrated trust can lead to unsafe overrides or uncritical acceptance of erroneous outputs. However, existing trust measurement approaches are often static, rely on subjective self-reports, and struggle to capture dynamic fluctuations during interactions. This study introduces a quantum-inspired computational framework for real-time trust estimation in human-AI teamwork. The model represents trust as a probabilistic state that evolves in response to multiple interaction variables (e.g., sentiment, task progress, response clarity, and accuracy proxies), enabling context-sensitive assessment without assuming fixed linear relationships. We instantiated the framework in a text-based interaction setting and compared model-estimated trust with 349 self-reported trust ratings from 15 participants. The quantum-based estimates closely tracked human trust ratings, with a mean bias of -0.78 (on a 0-100 scale) and a strong correspondence with self-reported trust ratings. This work demonstrates the feasibility of applying quantum-inspired modeling to real-time trust assessment for human-AI interaction. The framework provides a proof of concept for AI systems that monitor and adapt to evolving user trust in domains such as autonomous driving, clinical decision support, and other high-stakes environments.

Keywords: Trust; Quantum cognition; Human-AI interaction; Conversational AI; Trust dynamics

1. INTRODUCTION

1.1. Modeling Trust in Human-AI Interactions

Artificial intelligence (AI) systems are increasingly deployed in high-stakes sectors such as autonomous driving, healthcare diagnostics, and financial decision-making. As AI becomes more autonomous and integral to human workflows, its role shifts from tool to teammate, where AI agents act as semi-independent decision-making entities [1], [2]. In such collaborative environments, trust plays a critical role in determining system utilization and joint performance [3]. Miscalibrated trust, whether distrust or overtrust, can degrade decision quality and increase safety risks [4]. For example, low trust in automation may lead operators to override safe maneuvers, whereas excessive trust in decision aids can result in uncritical acceptance of faulty recommendations.

Beyond safety, inappropriate trust can undermine efficiency, transparency, and long-term adoption. In financial decision-making, overreliance on algorithmic trading may cause portfolio managers to miss emerging risks not modeled by the system, whereas excessive skepticism can lead them to discard valuable predictive signals. In defense and security applications, rapid swings in trust can disrupt coordinated operations when humans and AI agents must act under time pressure. These examples underscore that trust is not only a psychological construct but also an operational parameter that shapes system performance and resilience.

In this work, we focus on computational modeling of trust as a cognitive state that evolves during human-AI interactions. Rather than proposing a new questionnaire or static framework, we develop and evaluate a real-time trust estimator that ingests interaction signals and produces a moment-by-moment trust trajectory. Our goal is to support cognitive engineering and decision-

making by providing a generalizable modeling approach that could be embedded in AI systems to monitor and respond to human trust in any situation.

1.2.Challenges in Trust Modeling

Decades of work have produced a rich set of conceptual and mathematical models of trust in automation. Early efforts, such as the linear regression-based “trust transfer function” [5], offered foundational insights but captured only a portion of observed variance and treated trust as a relatively smooth, linear function of performance history. More cognitively grounded models, including those based on decision field theory incorporate stochastic evidence accumulation but are typically tuned for early-stage trust or specific laboratory tasks rather than continuous, real-time interactions [6], [7], [3].

In response, the human-automation and human-robot interaction communities have developed modern continuous and probabilistic models of trust, which we explicitly build upon rather than replace. [8] proposed a linear state-space model that treats trust as a latent state updated by performance and error signals over time. [9]’s OPTIMo model represent trust as a latent variable in a dynamic Bayesian network, inferring moment-to-moment trust from observed outcomes and operator interventions in supervisory human-robot collaborations. [10] integrated trust into a trust-POMDP, allowing a robot to infer trust, determining how its actions will change trust, and selecting actions that jointly optimize task performance and trust. More recent work has used clustering and state-space models to capture individual differences in trust dynamics [11] and nonlinear simulation models to explore complex trust trajectories under varying reliability profiles [12].

These models are comprehensively reviewed by [13], who summarize their strengths and limitations. In particular, continuous and Bayesian models currently provide time-varying trust

estimates and are compatible with control and planning architectures. Moreover, they can represent uncertainty over trust via latent variables and probabilistic state transitions. However, they typically assume the following: (1) Predefined functional forms and parameters. For example, [8]’s model uses a linear third-order state–space equation, and OPTIMo specifies conditional probability tables and transition probabilities a priori (or from offline data) [9]. These structures can approximate many trajectories, but they must be manually specified or reidentified when moving between tasks, interfaces, or user populations, limiting generalizability [13]. (2) Markovian and commutative updates. Most models assume that trust at time t depends on a summary of past evidence (e.g., current state and last observation). This implies that, conditional on the current state, the order of past events does not matter. However, empirical work shows that trust adjustments depend strongly on sequence and context. For example, failures often have a greater impact than successes do, and the same failure can be interpreted differently depending on whether it occurs early or late in the interaction [14], [15]. A trust-POMDP, for example, can encode different transition probabilities for “success after long success” versus “success after failure,” but doing so requires handcrafting additional states or features for many possible histories and quickly becomes unwieldy. (3) Relatively simple mappings from cues to trust. Continuous and Bayesian models often rely on a fixed set of observable predictors (e.g., reliability, intervention frequency, task performance) [8], [9]. However, recent work highlights that emotional tone, etiquette, task criticality, and outcome framing shape how the same performance event affects trust [16], [17], [18]. For example, an identical automation failure can produce very different trust adjustments depending on whether it occurs in a high-risk vs. low-risk context, whether the system apologizes, and whether the human ultimately corrects the error [14], [19]. Capturing such effects in classical models often requires adding new variables and interaction terms case-by-case.

Together, these limitations mean that current models are very good at representing smooth, performance-driven trust dynamics but are less well suited to three phenomena that are increasingly documented in the literature: (1) Nonlinear, asymmetric updates. Trust can decrease sharply after a salient failure and increase more modestly after successes, especially when outcomes are undesirable or when humans attribute responsibility differently [14], [15]. Encoding such asymmetry in linear state-space models or DBNs requires ad hoc nonlinearities or context-specific parameter changes. (2) Contextuality and interdependence. Trust is intertwined with reliance, self-confidence, affect, and perceived risk [16], [20], [21]. For example, negative effects or low self-confidence can change how a user interprets the same reliability information [17]. Classical models typically treat these as separate covariates or layers, which makes it difficult to represent situations where “performance + affect + risk” interact in nonadditive ways. (3) Order and interference effects. Decision science and quantum cognition studies show robust order effects, where answering questions or receiving evidence in different sequences leads to systematically different judgments [22], [23], [24]. Similar patterns appear in trust dynamics: e.g., a failure after a long streak of successes may be discounted as an anomaly, whereas the same failure early on can trigger sharp distrust and slower recovery [14], [25]. Capturing this “history-dependent meaning” of events is challenging for models that rely on Markovian assumptions and commutative updates.

Therefore, we do not claim that classical models cannot handle continuity or uncertainty; rather, our goal is to explore a modeling approach that is better aligned with context dependent, noncommutative, and interdependent aspects of trust, which have been highlighted as open challenges in recent surveys of mathematical trust models [13].

1.3. Quantum Mechanics and Modeling

To address these challenges, we draw on tools from quantum probability and quantum-inspired computation as a complementary modeling paradigm. Quantum cognition research has shown that quantum probability theory can naturally account for order effects, contextuality, and interference in human judgment and decision making—cases where classical probability and Bayesian models struggle unless they are heavily augmented [22], [23], [24].

In our approach, trust is treated as a latent state vector in a high-dimensional Hilbert space. A qubit (or set of qubits) encodes a superposition of “low-trust” and “high-trust” states, reflecting the idea that trust is continuous and may be partially indeterminate until it is expressed in behavior or self-reports [26]. Additional qubits represent related factors such as perceived performance, sentiment, and task criticality. Entanglement allows these variables to be represented jointly in a nonseparable way so that a change in, for example, perceived risk can alter the structure of the trust state rather than merely adding a constant offset. Crucially for our purposes: (1) Noncommutative updates. In a quantum framework, applying operator A then B generally yields a different state than applying B then A . This directly encodes order effects: a failure after successes can be represented by a different sequence of operators than a success after failures, without manually enumerating all possible histories. (2) Interference between cues. When multiple qubits are entangled, the combined state can exhibit interference terms. These can be used to model situations where the joint effect of performance and sentiment on trust is not simply additive. For example, when a polite explanation significantly dampens the negative impact of a minor failure or when negative effect amplifies the impact of a moderate error. (3) Contextual measurement. Measuring the state along different “bases” (e.g., asking for a self-reported trust rating vs. observing reliance behavior) can yield different projections of the underlying state. This aligns

with the observation that self-reported trust and behavioral measures sometimes diverge, even when they are driven by the same underlying cognitive state [17], [27]. Thus, in this work, we use quantum probability as a formal language for representing the kinds of contextual, order-sensitive, and interdependent trust dynamics that are difficult to express compactly in classical state-space or Bayesian models. A preliminary step in this direction is [26]’s quantum random walk model of trust, which uses superposition over a multidimensional Hilbert space to represent trust dynamics at a purely mathematical level. Our work extends this line of inquiry by (a) integrating a quantum-inspired model with real-time human-AI interaction, (b) incorporating multiple interaction-derived variables (e.g., sentiment, clarity, and performance proxies) into an entangled state representation, and (c) empirically comparing the model’s trust estimates with human self-reports over time. This provides a more direct test of whether a quantum-inspired representation can serve as a practical trust estimator within human-AI systems rather than remaining a purely theoretical construct.

1.4. Research Objective

To address the above gaps, the objective of this study is to design and empirically evaluate an automated, quantum-inspired model capable of estimating trust in real time during human-AI interactions. The model represents trust as a quantum state that evolves with each exchange and is influenced by multiple variables derived from the interaction (e.g., perceived system performance, user sentiment, and task context).

We embed this model in a conversational AI setting and use it to generate trust estimates after each interaction turn. These estimates are then compared with participants’ self-reported trust ratings collected during the session. This design allows us to examine whether the quantum-inspired estimator (a) produces trust levels that are comparable to those of human self-reports and (b) tracks the trend of trust over time as interaction unfolds. Our focus is on evaluating the

feasibility and calibration of the quantum estimator rather than on exhaustively modeling all factors that influence trust in every domain.

2. METHOD

The quantum-based model developed in this research is designed to track how trust shifts over time during human-AI interactions. The model estimates trust after each conversational turn, t , and visualizes it as a real-time line graph, enabling analysts to observe how specific AI responses affect the user's trust trajectory.

Because many human-AI interactions are text-based, this implementation relies solely on the text exchanged between the human and the AI. The front-end interface is a web-based application where users can chat with the AI and simultaneously view the evolving trust graph. The AI is powered by OpenAI's GPT-4o-mini model, which is accessed through the OpenAI API so that the interaction quality is comparable to that of standard large language model interfaces. The front end is implemented via Streamlit, whereas the back end runs the quantum-based trust estimator built in Python with the Qiskit framework.

In line with the goals of cognitive engineering and decision making, this implementation should be viewed as a general trust-estimation framework instantiated for a text-only conversational setting rather than a model tied to a specific task or domain.

2.1. IBM Quantum Computing Framework - Qiskit

To implement the trust model, a framework was needed that could represent quantum states and operations while still allowing access to the underlying linear-algebraic structure (state vectors, unitary operators, probability amplitudes). Qiskit, IBM's open-source quantum computing framework, was selected for this purpose. It provides tools for defining quantum circuits, applying

gates, and extracting measurement probabilities, which are used here to represent a latent trust state and its evolution over time.

In our model, a multiqubit quantum state encodes trust and its influencing factors. The trust qubit represents the user’s current trust level, whereas additional qubits represent variables such as sentiment and reliance. Qiskit’s state vector simulation is used to evolve this joint state and to compute the probability that the trust qubit collapses to $|1\rangle$; this probability is then mapped to a 0-100 trust scale. Three gate families form the core of the implementation: (1) Hadamard (H) gate: The Hadamard gate is used to initialize the trust qubit into a superposition state, reflecting the idea that trust at the start of a new interaction is uncertain rather than purely “low” or “high.” When applied to $|0\rangle$, it yields an equal superposition of $|0\rangle$ and $|1\rangle$ [28]. (2) Rotation gates (RY): Rotations around the y-axis are used to adjust the state of individual qubits on the basis of numerical inputs such as initial trust, perceived risk, sentiment, or proxy measures of reliance [29]. The angle θ affects the magnitude and direction of the update on the Bloch sphere. (3) Controlled RY (CRY) gate: To couple trust with its influencing factors, we use controlled rotations, where the state of one qubit modulates a rotation on another. The CRY gate applies an $R_y(\theta)$ rotation to a target qubit only when the control qubit is in $|1\rangle$, which aligns with the idea of entanglement introduced in Section 1.3 above. The CRY gate is used to encode dependencies such as “if sentiment is negative, shift the trust state toward lower values.” [30]. We deliberately use CRY rather than more extreme gates (e.g., CNOT) to allow graded rather than binary changes.

Together, these operations allow trust and its influencing factors to be represented as a multiqubit state that can be initialized, updated, and measured in a principled way while still being tractable for simulation on classical hardware.

2.2.Influencing Factors

Measuring trust in human-AI interactions necessitates careful attention to various temporally dynamic and static factors that influence the user's trust state throughout an interaction. Building on the framework proposed by [3], this study incorporates only those trust-influencing factors that are measurable through text-based communication, thus avoiding the need for multimodal data collection. The selected factors are categorized into two types: preinteraction-based and interaction-based. Preinteraction factors are essential for establishing an initial baseline and include the initial trust state, the perceived level of risk, and the user's prior knowledge or confidence. The initial trust state reflects the user's predisposed attitude toward the AI system, which can be informed by previous experiences or a general disposition toward automation. Studies have shown that this predisposition can significantly affect subsequent reliance behavior and trust dynamics [17]. The perceived risk, or vulnerability, in a situation plays a moderating role between the willingness to trust and actual trusting behavior: users are more likely to withhold trust when the stakes are high, and a mistake by the system could result in serious consequences [3]. In parallel, the user's prior domain knowledge or confidence in the subject matter alters how critically they evaluate the AI's responses; individuals with greater expertise are more likely to scrutinize outputs, whereas those lacking knowledge may place uncritical trust in the system [16]. As these factors are unaffected by feedback from the interaction and do not vary temporally, they are measured prior to engagement. In contrast, interaction-based factors evolve over time and are monitored continuously during the user's engagement with AI. Sentiment, assessed through natural language processing techniques, captures the emotional tone conveyed in the user's text [31]. Sentiment polarity (positive, negative, or neutral) can indicate changes in trust, with negative sentiment often correlating with a decline in perceived system reliability. Reliance on the AI

system is indirectly inferred through three measurable indicators: the task completion rate, the clarity of the AI's response, and its perceived accuracy. A low task completion rate may signal difficulty in achieving goals with AI, thus eroding trust. Similarly, if the AI output lacks clarity, users may experience frustration or disengagement, opting instead to rely on their own knowledge. Finally, perceived accuracy, i.e., whether the system's information aligns with factual correctness or user expectations, strongly influences ongoing trust. Inaccuracies or contradictions can lead to rapid trust degradation, particularly in contexts involving high informational risk. These interaction-based features are critical for real-time trust modeling and adaptation. A visual depiction of these factors and their temporal relationships with respect to trust is provided in Figure 1.

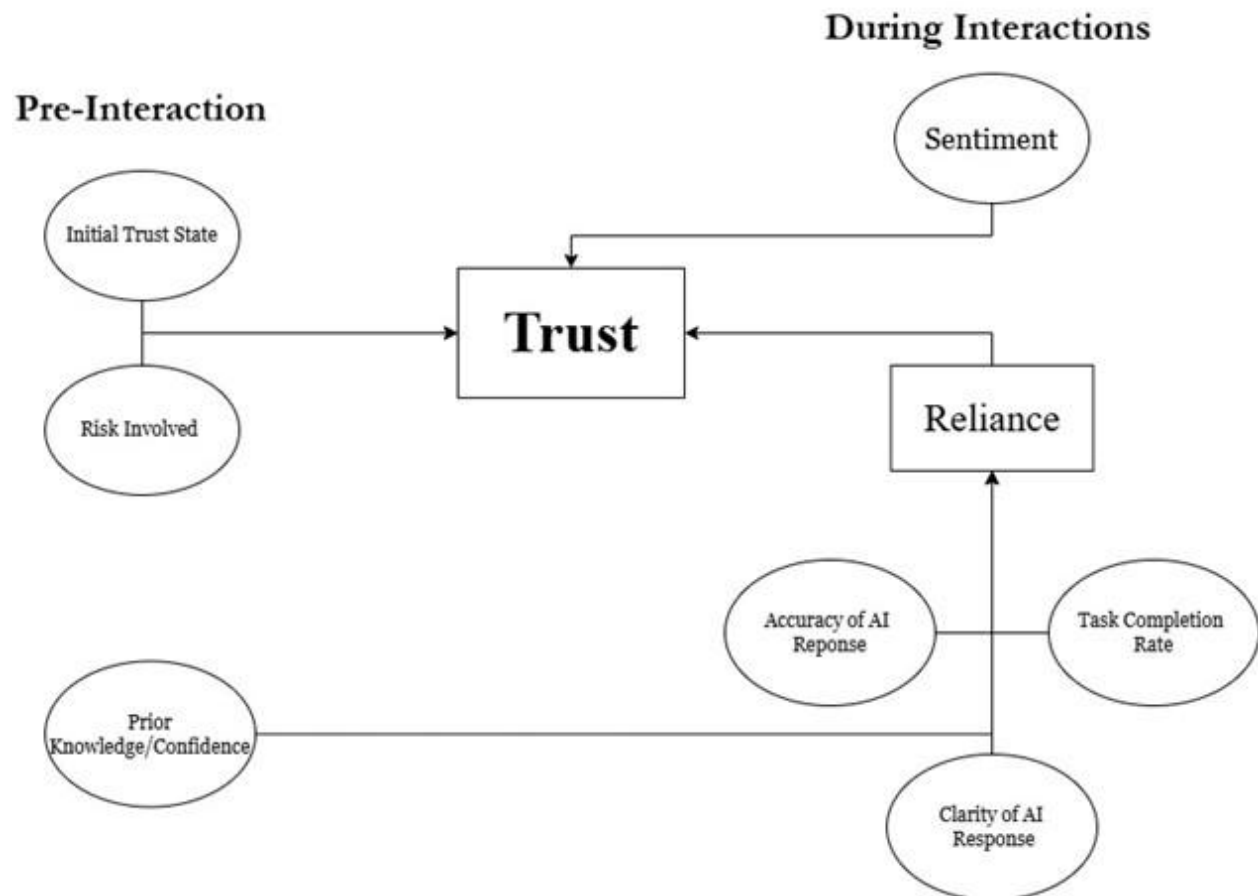


Figure 1: Influencing factors

All simulations use state vector evolution (no sampling noise) and compute trust as the marginal probability that the trust qubit (qubit 0) is in the computational basis state $|1\rangle$ after each conversational turn. Let $|\psi_t\rangle$ be the n -qubit state vector after applying the full gate sequence for turn t . Trust is computed as:

$$\text{Trust}_t = 100 \cdot P_t(q_0 = 1) \quad (1)$$

where the marginal probability is obtained by summing the amplitudes over all other qubits:

$$P_t(q_0 = 1) = \sum_{b \in \{0,1\}^{n-1}} |\langle 1b | \psi_t \rangle|^2. \quad (2)$$

This probability is computed directly from the statevector without physically measuring (collapsing) the full register, so the model remains a deterministic estimator given the extracted text feature. To ensure full reproducibility, the model uses a fixed qubit register, as shown in Table 1.

Table 1: Qubit register and interpreted constructs

Qubit index	Name (construct)	Type	When updated	Observable used at output
0	Trust	Latent state	Initialized preinteraction; updated every turn	$P(q_0 = 1) \rightarrow \text{Trust}_t$
1	Sentiment	Interaction-based	Every turn from user text sentiment score	Indirect via entanglement to Trust
2	Reliance	Interaction-based	Initialized preinteraction; updated every turn	Indirect via entanglement to Trust
3	TCR proxy	Interaction-based	Each turn after baseline threshold	Indirect via entanglement to Reliance
4	Clarity	Interaction-based	Every turn from readability of AI response	Indirect via entanglement to Reliance

5	Accuracy proxy	Interaction-based	Every turn from AI self-confidence score	Indirect via entanglement to Reliance
---	----------------	-------------------	--	---------------------------------------

2.3.Trust

For the purposes of this study, **trust** is defined as *the extent to which a person is confident in and willing to act on the basis of the recommendations, actions, and decisions of an artificial agent* [32]. This definition motivates the choice of factors that influence the trust qubit. In constructing the quantum model, trust is represented by a single qubit (qubit 0), which is initially placed in a superposed state via the Hadamard gate. As discussed previously, modeling trust in superposition allows it to exist in a coherent blend of possible outcomes, reflecting uncertainty, until it is observed through measurement. Following this initialization, the trust state is influenced by two user-specific parameters: initial trust and perceived risk. These parameters are elicited at the outset of interaction via self-reports on a scale from 0 to 10, where 0 corresponds to an absence of trust or risk, and 10 denotes complete trust or extreme perceived risk, respectively. To translate the initial trust value into a quantum rotation, the following formula is employed:

$$\theta_{initialTrust} = ((\beta - 5) * 0.2) * \frac{\pi}{2} \quad (3)$$

Here, β denotes the user's self-reported level of trust. Constant 5 offsets the 0-10 trust scale so that $\beta = 5$ corresponds to neutrality, placing the trust qubit at the Bloch equator (i.e., equal probability of collapsing to 0 or 1). A scaling factor of 0.2 was established through an iterative calibration process using a separate pilot dataset (N=5) distinct from the validation cohort reported in this study. This separation ensures that the parameters are not overfit to the test participants, maintaining the integrity of the subsequent validation analysis. Multiplication by $\pi/2$ then converts this normalized span into a physically meaningful angular displacement: since $|0\rangle$ and $|1\rangle$ are

separated by π on the Bloch sphere, $\pm\pi/2$ corresponds to a half-rotation that moves the state from the equator toward either pole. This formulation ensures that values below 5 rotate the qubit toward $|0\rangle$ (reduced trust), whereas values above 5 rotate it toward $|1\rangle$ (increased trust), which is consistent with the desired trust modulation scheme. Second, for the perceived risk, the rotation angle is calculated using:

$$\theta_{risk} = \left(-\frac{risk}{10}\right) * \frac{\pi}{8} \quad (4)$$

where *risk* is the user’s self-reported perceived risk (0-10). The division by 10 normalizes the 0-10 input scale, whereas the smaller factor $\pi/8$ was chosen through pilot tuning to keep risk a secondary influence relative to initial trust, preventing it from overpowering the state dynamics. In practice, $\frac{\pi}{8}$ was the smallest value that still produced visibly different trajectories for low- versus high-risk scenarios while keeping risk-induced changes clearly secondary to initial trust in magnitude. The negative sign captures the inverse relationship between risk and trust, i.e., higher perceived vulnerability reduces the degree of trust. Once both the trust and risk rotations are applied via RY gates, the trust qubit is initialized. As the interaction progresses, trust is directly affected by two factors, sentiment and reliance. Both of these temporally dependent factors are modeled as their own qubits (qubits 1 and 2, respectively), as they change with each interaction. Sentiment is currently quantified via OpenAI’s gpt-4o-mini model. Sentiment scoring uses a fixed prompt that returns a single scalar $\alpha_t \in [-1,1]$ from the user’s text at turn t . The API settings are held constant across runs. A score of 1 indicates strongly positive sentiment toward the AI, whereas -1 reflects strong negativity. Prior work has demonstrated that GPT-4o-family models can be effectively applied to sentiment analysis tasks when prompted in this manner [33]. This sentiment score is used to rotate the sentiment qubit about the y-axis according to the expression:

$$\theta_{sentiment,t} = \alpha_t \cdot \frac{\pi - S_{t-1}}{\delta_t} \quad (5)$$

Here, α_t is the sentiment score at turn t , S_{t-1} is the cumulative sentiment rotation applied up to turn $t - 1$, and δ_t is the dilution factor for turn t . The dilution factor δ_t increases deterministically each turn to dampen updates, with $\delta_t=1$ and $\delta_{t+1} = \delta_t + 0.5$. The increment of 0.5 was set through iterative testing to dampen sudden swings and model the gradual stabilization of trust over time. Specifically, δ_t was constrained so that early interactions (small t) could still meaningfully shift the trust state, whereas later interactions (larger t) produced diminishing marginal updates, reflecting empirical findings that trust changes rapidly at first and then stabilizes with experience [34], [35]. This factor models the empirical observation that trust tends to stabilize over repeated interactions, thus dampening future updates [35]. After the sentiment qubit has been updated, it is entangled with the trust qubit via the CRY gate. The angle of rotation for the CRY gate is dependent on the polarity of the sentiment:

$$\theta_{(cry,sent,t)} = \begin{cases} \frac{(0 - \gamma_t)}{\delta_t}, & \alpha_t < 0 \\ \frac{(\pi - \gamma_t)}{\delta_t}, & \alpha_t > 0 \end{cases} \quad (6)$$

Here, γ_t is the current angle of the trust qubit at turn t , anchoring the adjustment to its present state. The reference points 0 and π represent the poles $|0\rangle$ and $|1\rangle$ on the Bloch sphere, which are chosen to model the opposing pull of negative vs. positive sentiment. Division by the dilution factor δ ensures that this influence scales consistently with the stabilization process described above.

Reliance is handled analogously and is described in the following section. In combination, these mechanisms allow both transient events (e.g., sharp negative sentiment) and cumulative experience (via the growing δ) to shape the trust trajectory.

2.4. Reliance

In the context of human-AI interactions, reliance is defined as observable behavior in which users either follow or override AI-generated recommendations [36]. Within our quantum-based trust modeling framework, reliance is represented as an independent qubit (qubit 2) to reflect its dynamic, temporally evolving nature. Before interactions, the qubit is rotated around the y-axis via the confidence score that the user provides from a scale of 0-10 in a similar fashion to the perceived risk. Thus, prior knowledge/confidence ($conf$, 0-10) initializes the reliance qubit via:

$$\theta_{conf} = \left(-\frac{conf}{10}\right) * \frac{\pi}{8} \quad (7)$$

Owing to the constraints of a text-only system, direct measurement of reliance is infeasible. Instead, we operationalize reliance through three quantifiable proxies, namely, the task completion rate (TCR), clarity, and accuracy, each of which is modeled as a distinct qubit (qubits 3, 4, and 5, respectively).

To account for reading time and system latency, the TCR is approximated via the number of follow-up questions the user asks. After the first two queries, the TCR begins to influence reliance. The TCR qubit is rotated by:

$$\theta_{TCR,t} = -((\varepsilon_t - 2) * 0.1) \cdot \frac{\pi - C_{t-1}}{\delta_t}, \quad \varepsilon_t > 2 \quad (8)$$

Here, ε_t denotes the number of follow-up questions, with the subtraction of 2 setting a neutral baseline (i.e., a small number of clarifying questions are expected and not penalized). C_{t-1} is the cumulative TCR rotation applied up to turn $t - 1$, whereas δ_t is the interaction-based dilution factor described earlier. The coefficient 0.1 was calibrated to attenuate the effect of the TCR, since task duration may also reflect user behavior or question complexity rather than a lack of trust in the AI. The value 0.1 was chosen as the largest coefficient that kept TCR-induced changes to rely

within a bounded range over the observed number of turns, ensuring that even extreme values of ε did not cause abrupt, implausible shifts in the reliance qubit. After rotation, the TCR qubit is entangled with the reliance qubit via the following angle:

$$\theta_{cry} = \frac{(0 - \rho_t)}{\delta_t} \quad (9)$$

where ρ_t is the current angle of the reliance qubit at interaction t . Reference point 0 ensures that the entanglement aligns the reliance qubit toward the $|0\rangle$ state when the TCR indicates reduced confidence. In parallel, clarity is integrated into the model by quantifying the readability of AI-generated responses via the TextStat Python library. TextStat computes multiple readability indices, including the Flesch Reading Ease, Flesch-Kincaid Grade, Gunning Fog, SMOG, Automated Readability Index, and Dale-Chall. Each index is normalized to $[0, 1]$ with the direction aligned so that higher values indicate clearer/easier text and the clarity score, $clar_t$, is the mean of the normalized indices. This normalized score then drives a rotation of the clarity qubit about the y-axis, which is subsequently entangled with the reliance qubit via the same entanglement mechanism as above.

Finally, the AI’s self-assessed confidence in its own response is quantified, producing a confidence score ranging from -1 (entirely inaccurate) to 1 (highly accurate), with this score representing the model’s internal confidence. This score is then applied to the accuracy qubit via a methodology analogous to that used for clarity and sentiment. The accuracy qubit is then entangled with the reliance qubit to update the reliance state. After each turn t , the updated reliance qubit is itself entangled with the trust qubit. This allows behavioral and performance-related cues (TCR, clarity, accuracy) to influence the latent trust state over time, which is consistent with the idea that trust and reliance are related but distinct constructs. A summary of all the steps is provided in the algorithm described below:

393 **Algorithm 1. Per-turn quantum trust update (state vector estimator)**

394 Inputs (pre-interaction): $\beta \in [0,10]$, risk $\in [0,10]$, conf $\in [0,10]$

395 Inputs (each turn t): user text U_t , AI response text A_t

- 396 1. Initialize dilution: set $\delta_1 = 1$.
- 397 2. Initialize the circuit/state: Prepare $n = 6$ qubits in $|0 \dots 0\rangle$.
- 398 3. Trust initialization: Apply $H(q_0)$; apply $R_y(\theta_{\text{initTrust}})$ to q_0 ; apply $R_y(\theta_{\text{risk}})$ to q_0 .
- 399 4. Reliance initialization: apply $R_y(\theta_{\text{conf} \rightarrow \text{rel}})$ to q_2 .
- 400 5. For each turn $t = 1, 2, \dots, T$:
 - 401 ○ Feature extraction: Compute $\alpha_t = \text{Sentiment}(U_t) \in [-1,1]$; compute $\kappa_t =$
 - 402 Accuracy Confidence(A_t) $\in [-1,1]$; compute $\text{clar}_t = \text{Clarity}(A_t) \in [0,1]$; update
 - 403 the follow-up counter ε_t .
 - 404 ○ Sentiment update: Apply $R_y(\theta_{\text{sentiment}}(\alpha_t, \delta_t, \phi_t))$ to q_1 ; then, apply
 - 405 $CR_y(\theta_{\text{cry,sent}}(\alpha_t, \delta_t, \gamma_t))$ with control $q_1 \rightarrow$ target q_0 .
 - 406 ○ Reliance proxies: if $\varepsilon_t > 2$, apply $R_y(\theta_{\text{TCR}}(\varepsilon_t, \delta_t, \sigma_t))$ to q_3 then
 - 407 $CR_y(\theta_{\text{cry,TCR}}(\delta_t, \rho_t))$ with $q_3 \rightarrow q_2$. Clarity $R_y(\theta_{\text{clar}}(\text{clar}_t))$ is applied to q_4 , and
 - 408 then $CR_y(\theta_{\text{cry,clar}}(\delta_t, \rho_t))$ with $q_4 \rightarrow q_2$. The accuracy $R_y(\theta_{\text{acc}}(\kappa_t))$ is applied to
 - 409 q_5 , and then $CR_y(\theta_{\text{cry,acc}}(\delta_t, \rho_t))$ with $q_5 \rightarrow q_2$.
 - 410 ○ Reliance $\rightarrow$ Trust: apply $CR_y(\theta_{\text{cry,rel}}(\delta_t, \gamma_t, \rho_t))$ with $q_2 \rightarrow q_0$.
 - 411 ○ Output trust: computes $\text{Trust}_t = 100 \cdot P_t(q_0 = 1)$ from the state vector.
 - 412 ○ Update dilution: set $\delta_{t+1} = \delta_t + 0.5$.

413

414 The fixed symbol definitions used throughout the model are provided in Supplementary

415 Table S1, and a summarized per-turn gate and update schedule is given in Supplementary Table

S2. Unless otherwise stated, all rotations follow the Qiskit convention $R_y(\theta)$, with θ expressed in radians and measurements performed in the computational basis.

2.5.LLM-based feature extraction

Sentiment and related conversational features were extracted using OpenAI's *gpt-4o-mini* model via the Chat Completions API. The model was prompted using a fixed system instruction designed to elicit a scalar sentiment score in the range $[-1,1]$ from the user's text at each interaction turn. The prompt format, model identifier, and API parameters were held constant across all experimental runs, and no adaptive prompting or fine-tuning was performed during the study.

Model outputs were constrained to numeric values only and were parsed directly from the model's response. To avoid information leakage, sentiment scores were computed solely from the user's textual input at each turn and were not conditioned on self-reported trust ratings, experimental condition labels, or prior trust estimates. While large language model outputs may vary across model versions or over time, the use of a fixed prompt and fixed API configuration ensured consistent feature extraction within the present study. The full prompts and implementation details are provided in the Supplementary Information.

2.6.How can the quantum model be used?

After successfully installing and running the program ([GitHub](#)), a website is presented to users, as shown in Figure 2. This webpage contains three sliders, each allowing the user to input scores from 0--10 for initial trust, perceived risk and prior knowledge.

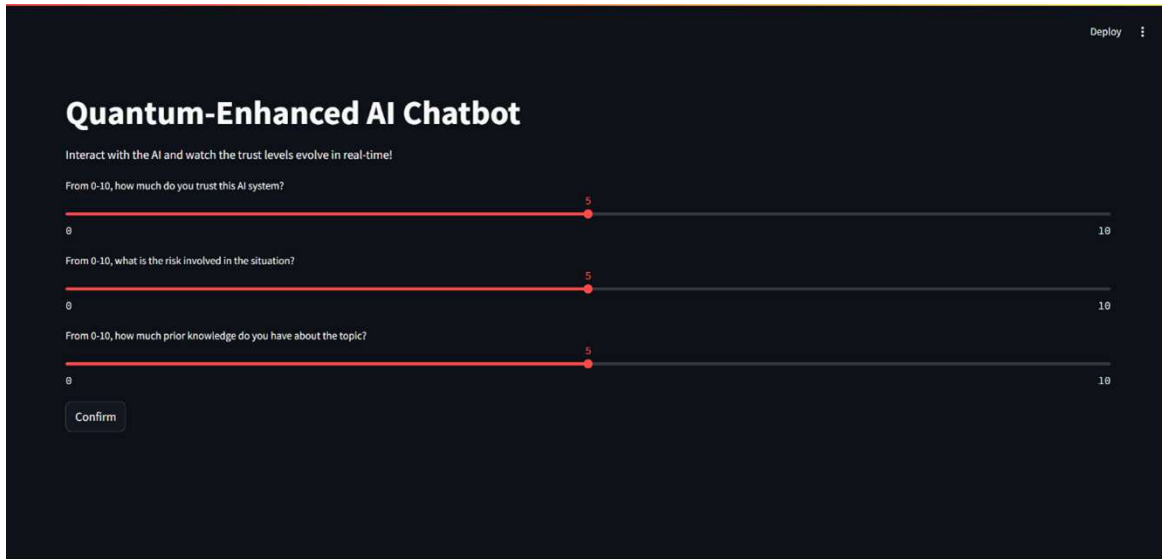


Figure 2: Starting interface and sliders

After the scores have been added using the sliders, the users can click “Confirm”, which will take them to another page, as shown in Figure 3.

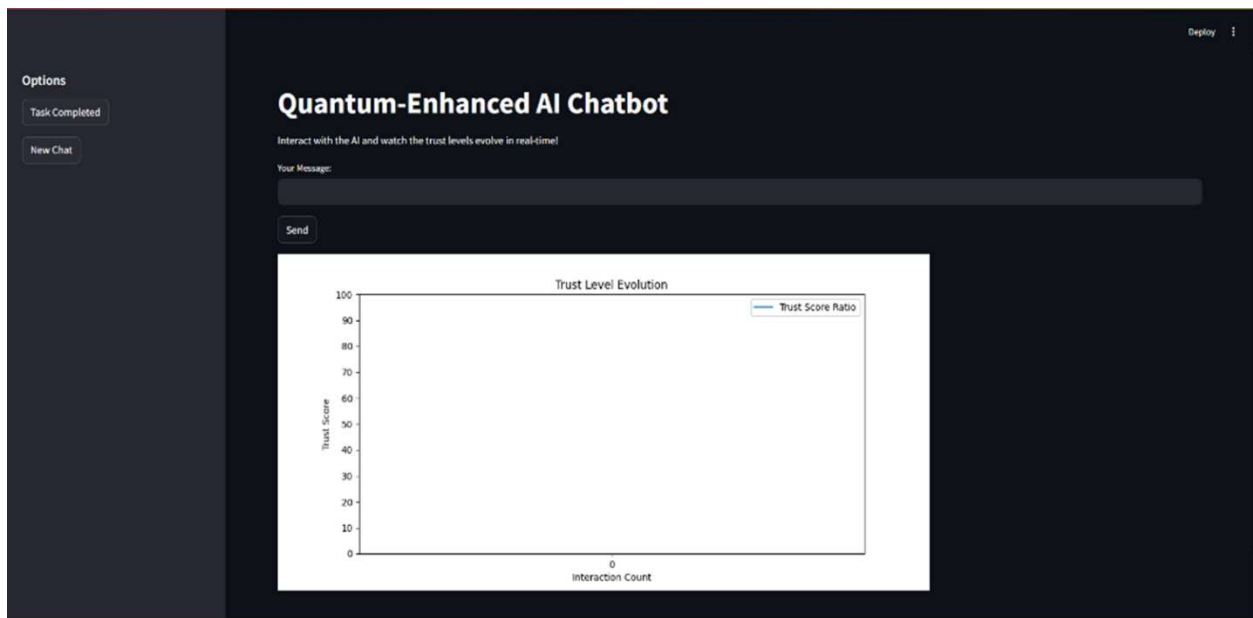


Figure 3: Interaction Page

This is the main interface of the model, where users are able to interact with an inbuilt gpt-4o-mini model via the text space provided below “Your Message”. Once the users have

entered their message in the space provided, they can press “Send”, which will generate and display a response from the gpt model. Crucially, the quantum model receives only the text interaction data and the initial slider values; it is blinded to the user's intermediate and final trust ratings to prevent data leakage.

Below the response, a line graph will be updated; this graph will show the results of the quantum model, which live to the viewer. For the validation, this graph was removed to ensure that no biases were incurred. An example of such a graph is displayed in Figure 4.

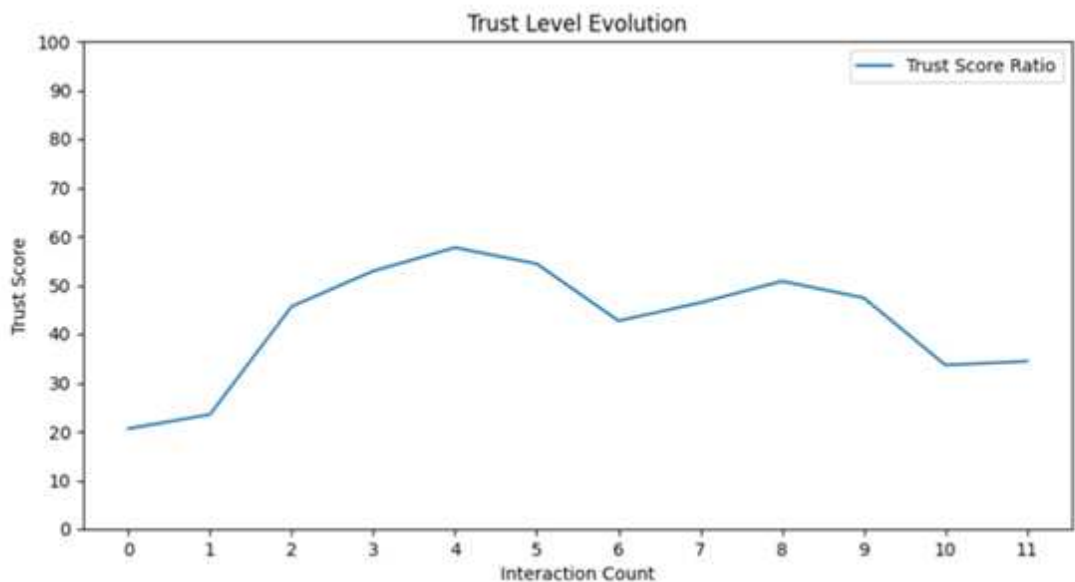


Figure 4: Real-time trust graph

Once the user has completed their current task, they can click the “Task Completed” button on the side panel, which will complete the task on the backend and store the current user’s final trust value. The user can then generate a new chat if they want to. All model parameters, gate mappings, and analysis scripts are available upon reasonable request and will be released publicly upon publication.

2.7. Model Validation

A study was conducted with 15 individuals (3 females, 12 males; age: $M = 18.47$ years, $SD = 1.30$ years), each of whom was permitted up to 6-8 interactions with the AI model. The study was executed through an online meeting platform, Microsoft Teams, wherein the participants were given access to the AI model for interactions, and the scenarios were recorded. The study protocol was approved by the University of Calgary Conjoint Faculties Research Ethics Board (REB25-0370). All participants provided informed consent prior to participation. All methods were performed in accordance with the relevant guidelines and regulations of the University of Calgary Conjoint Faculties Research Ethics Board and in compliance with applicable institutional and international research standards. In this section, we outline participant demographics and the study procedures.

All the participants had prior experience using GPT-4o-mini-based AI systems. Each had completed at least a high school education and demonstrated proficiency in using computers in English. All the sessions were conducted in a quiet, controlled environment to reduce distractions and ensure consistent interaction conditions across participants.

To validate the quantum trust model, participants interacted with the integrated GPT-4o-mini system across two distinct task scenarios: one low-risk scenario and one high-risk scenario. Since the perceived risk of a situation is a major contributor to the change in initial trust [3], we wanted to test performance in both situations. After each interaction with the AI system, the participants were allowed to record their perceived level of trust on a scale of 0-100. To ensure consistency across responses, participants were instructed to base their trust ratings on how confident they felt in the AI recommendations and how willing they would be to act on them. Trust was conceptualized as a combination of perceived confidence in the AI output and the likelihood

of relying on it without further verification. These self-reported trust values were treated as the ground truth and compared with the model-generated trust scores for evaluation, with a focus on statistical analysis.

The low-risk scenario involved a task in which participants were asked to generate a healthy and budget-conscious (\$70 budget) weekly meal plan tailored to their dietary restrictions (e.g., vegan, gluten-free) and culinary preferences. The participants were encouraged to explore diverse meal options and could inquire about local grocery stores, discounts, and relevant nutritional information. This scenario was classified as low risk because it involved routine, everyday decision-making with minimal personal or financial consequences. Prior work suggests that such noncritical tasks typically elicit baseline or stable trust levels in AI, given the low perceived vulnerability involved [2].

In contrast, the high-risk scenario simulated an emergency context in which participants returned home to discover flooding in their apartment. They were tasked with seeking AI guidance on identifying the cause of the incident, ensuring personal safety, preserving belonging, and determining appropriate next steps. These included contacting landlords, arranging for repairs, handling documentation, and communicating with insurance providers. This scenario was selected to represent a high-risk situation due to the potential for significant financial loss, emotional stress, and the urgency of action. Previous articles highlight that user trust becomes more variable and fragile in high-risk or high-stakes contexts, especially where safety or livelihood is involved [3].

To ensure sufficient data for modeling dynamic trust states and enough comparisons between the ground truth (survey) and the model values, a minimum of five interactions ($n \geq 5$) was required for each scenario. The participants could exceed this threshold if they wished to do so. To allow for smoother interaction and fine-grained analysis of how trust evolved throughout

the conversation, the participants were instructed to provide only one new piece of information per prompt. The trust scores from all the conversations were logged for subsequent analysis.

2.8.Hypothesis

For the trust estimation model to be practically useful in a real-time human-AI team context, it must satisfy two essential conditions. First, the model's trust outputs need to be calibrated to the same numerical scale as the human trust ratings. Calibration ensures that the model's baseline trust values correspond with users' own trust assessments, which is necessary for setting adaptive thresholds (e.g., triggering an alert when trust drops below 40%). Without this alignment, the model's trust estimates could be skewed by unknown bias, leading to misinterpretation of fluctuations [37]. Once calibration was established, the next step was to evaluate the model's responsiveness i.e. its ability to track moment-by-moment shifts in user trust. Responsiveness is a critical property for adaptive systems and is necessary for enabling fluid collaboration between humans and AI agents in dynamic environments [3]. Importantly, calibration and responsiveness are statistically independent; a model may be unbiased but unresponsive or reactive yet systematically misaligned. As such, each property requires a distinct hypothesis test.

Accordingly, two hypotheses were formulated.

- **H₁ (Calibration).** *There will be no significant mean difference between the trust values generated by the quantum model and participants' self-reported trust ratings across both high- and low-risk task scenarios.*

- **H₂ (Responsiveness).** *The trust values generated by the quantum model significantly correlate with participants' self-reported trust levels across both high- and low-risk task scenarios.*

To evaluate H₁, paired-sample mean comparisons and Bland-Altman bias analysis was used to assess the agreement between the model predictions and human ratings. H₂ was tested via Pearson and Spearman correlations, along with repeated measures mixed modeling. This sequential approach ensured that the trust estimator was placed on the correct measurement scale before its capacity to track trust changes over time was analyzed.

2.9.Data analysis

The analytic strategy was designed to follow a stepwise protocol, which is consistent with established recommendations for method-comparison studies [37] and prior research on human-automation trust [3]. The data were reshaped to a long format in Python 3.10 (*pandas* v2.1) and screened for randomness of missing cells with Little's MCAR test. For H₁ (Calibration), interaction-level trust scores were averaged within each Participant $\times$ Risk $\times$ Source cell. Paired-sample *t* tests were then used to assess whether the quantum model's trust values differed significantly from the participants' self-reported ratings. Bland-Altman plots were generated to visualize the level of agreement and detect any consistent bias across the trust spectrum. For H₂ (Responsiveness), the associations between model-generated trust values and participants' self-reported trust ratings were evaluated via both Pearson and Spearman correlation coefficients. To determine whether these correlations varied across task risk conditions, Fisher's *r*-to-*z* transformation was applied.

Prior to all the parametric tests, the statistical assumptions were verified. These included checks for the normality of paired differences via the Shapiro-Wilk test [38] and for homogeneity

of variance via Levene's test [39]. To exploit all 349 ratings while accounting for repeated measures, a linear mixed-effects model with a random intercept for participants was fitted in R 4.3.2 via *lme4* [40]. The model included a random intercept for each participant to control for individual baseline trust levels. Residual diagnostics were performed via Q-Q plots and scale-location plots to ensure that the assumptions of normality and homoscedasticity were met. Cook's distance was used to identify any influential outliers. All reported statistical outcomes include effect sizes (Cohen's d , r^2) and 95% confidence intervals to support interpretability and reproducibility.

3. RESULTS

3.1.Descriptive Analysis and Diagnostics

The dataset comprised 349 valid interaction ratings from 15 participants (6-8 interactions, 2 risk types, and 2 sources per person, $M = 23.3$), all of whom were undergraduate students. Missing interaction cells (27.3%) were randomly distributed (Little's MCAR test, $\chi^2 = 11.4$, $df = 14$, $p = .65$) and were therefore pairwise deleted. A mean score was computed for every participant $\times$ risk $\times$ source combination, yielding 30 paired observations for subsequent analyses. Table 2 summarizes the aggregated means. Shapiro-Wilk tests indicated that the paired differences (Quantum-Self-Reported) were approximately normally distributed ($W = 0.95$, $p = .15$), and Levene's test revealed homogeneous variance across risk scenarios ($F_{1,28} = 0.56$, $p = .46$).

Table 2: Descriptive Statistics

Risk Scenario	Source	<i>M</i>	<i>SD</i>	<i>n</i>
High	Self-Reported	51	16.72	15
	Quantum model	49.03	17.41	15
Low	Self-Reported	75.14	12.01	15
	Quantum model	75.56	14.2	15

3.2.Hypothesis 1 - Calibration (mean-level agreement)

To ensure the robustness of the findings despite the sample size ($N=15$), we utilized a linear mixed-effects model to account for participant-level variance and Bland-Altman analysis to quantify systematic bias. The inputs driving the model, specifically 'accuracy' (operationalized as the LLM's internal confidence score ³) and 'task completion' (operationalized via explicit user confirmation ⁴), were generated independently of the users' self-reported trust ratings to prevent circularity. Although open-ended chats introduce variability, these automated proxies provide a consistent, external criterion for reliance that remains distinct from the ground-truth trust signals used for validation.

Calibration was tested with a paired-sample t test on the 30 quantum-self-reported pairs ($t(29)=-0.89, p=.381, d=-0.16, 95\% \text{ CI} = [-2.56, 1.01]$). The mean bias was -0.78 trust units ($SD = 4.77$). A Bland-Altman analysis yielded limits of agreement of -9.25 - 7.69 and a root-mean-square error (RMSE) of 4.76 . These results confirm H_1 : the quantum model shows no systematic over- or underprediction.

3.3.Hypothesis 2 - Responsiveness (covariation)

Pearson and Spearman correlations were used to quantify the dynamic correspondence between sources. Across all risk scenarios, a strong positive association was observed (Pearson $r = .975, 95\% \text{ CI} [.94, .99], p = 8.8 \times 10^{-20}$; Spearman $\rho = .944, p = 4.7 \times 10^{-15}$). Separate analyses per risk scenario produced similarly strong associations (Table 3). A Fisher r -to- z comparison indicated no significant difference between the two correlation coefficients ($z = 0.30, p = .76$). Therefore, H_2 is supported.

Table 3: Pearson correlations between quantum and self-reported trust by risk type

Risk Scenario	r	95% CI	p
High ($n = 15$)	0.954	[.83, .99]	3.4×10^{-8}
Low ($n = 15$)	0.964	[.88, .99]	7.3×10^{-9}

Figure 5 illustrates the relationship between participant-level mean quantum-model trust estimates and corresponding self-reported means. Blue squares represent low-risk trials and red circles represent high-risk trials, with the dashed 45° line indicating the identity relationship. Data points from both conditions are distributed closely around the identity line, indicating good agreement between model estimates and self-reported trust. No systematic separation by risk conditions is evident, consistent with the comparable correlation magnitudes reported in Table 3.

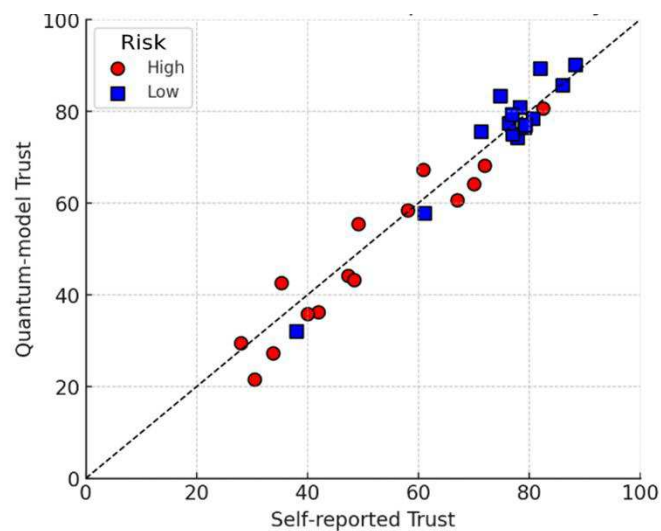


Figure 5: Trust data scatter plot

The calibration metrics showed no significant mean bias, and the covariation metrics demonstrated strong associations in both task contexts, supporting Hypotheses 1 and 2. The mixed-effects model corroborated these findings across all 349 individual ratings.

3.4. Additional Statistical Analysis Robustness and Equivalence Checks

In addition to hypothesis tests, we evaluated practical equivalence between the model and self-reports via two one-sided tests (TOSTs) on the aggregated Participant×Risk pairs [41]. With a prespecified margin of ± 5 trust units (5% of the 0-100 scale), both one-sided tests were significant ($t=4.84$ and $t=6.62$; $p<.001$), establishing equivalence rather than merely “no

difference.” A stricter ± 3 -unit sensitivity check also passed ($t=2.55$ and $t=4.33$; $p \leq .01$), indicating that any residual mean bias is practically negligible under even conservative thresholds relevant for UI triggers (e.g., adaptive warnings). A power analysis based on the observed variability of paired differences revealed that the study had $\approx 80\%$ power to detect mean calibration errors as small as ~ 2.4 units, underscoring that the null finding is not a byproduct of low sensitivity [41].

To prove proportional or systematic bias, we regressed quantum estimates on self-reports. An ordinary-least-squares fit yielded a slope of 1.07 (95% CI [0.97, 1.16]) and an intercept of -5.02 , so the identity line lies within the slope CI, with no evidence of proportional bias. Interpreting the line of best fit, the expected model-human difference is approximately -3.7 units at a low trust level (20) and $\sim +0.4$ units at a high trust level (80), crossing zero near 75, which indicates only small, directionally balanced deviations across the practical range. An errors-in-variables (Deming) regression ($\lambda=1$) produced a slope of 1.10 and an intercept of -6.90 , corroborating the absence of meaningful bias when measurement error in both sources is acknowledged [42], [43]. Finally, the uncertainty around the Bland-Altman limits is tight: the 95% confidence band around each limit is approximately ± 3 units, so even when accounting for sampling error, the agreement envelope remains well below 10% of the full scale—a tolerance typically acceptable for trust-thresholding in adaptive interfaces [44].

4. DISCUSSION

The present study provides empirical support for the feasibility of a quantum-inspired estimator for tracking human trust during conversational human-AI interactions. The results indicate that the estimator produced trust values that were well calibrated to participants’ self-reported trust, with no statistically significant mean bias. In addition, the model demonstrated

strong temporal correspondence with self-reported trust trajectories across interaction turns. Together, these findings suggest that the proposed framework can generate stable and responsive trust estimates on the same numerical scale used by human raters, satisfying key requirements for real-time trust monitoring in adaptive AI systems.

Beyond calibration and responsiveness, the results were consistent across task contexts that differed in perceived risk. Agreement between model estimates and self-reported trust was comparable in both low-risk and high-risk scenarios, and no significant source $\times$ risk interaction was observed. This stability contrasts with prior work reporting degradation of trust modeling performance under higher-stakes conditions, where variability and volatility tend to increase. One possible explanation is that the joint representation of contextual cues, such as sentiment and reliance, within an entangled state allows the model to accommodate risk-related variance without requiring explicit reparameterization. Importantly, this interpretation should be viewed as a plausible mechanism rather than a definitive causal explanation, given the scope of the present study.

The observed robustness across individuals further supports the generality of the framework. Although participants exhibited substantial between-subject variability in baseline trust levels, the mixed-effects analysis indicated that the model-maintained agreement with self-reports without requiring individualized parameter fitting. This suggests that the quantum-inspired representation may offer a flexible structure for modeling trust dynamics that generalize across users, at least over short interaction horizons.

From a theoretical perspective, the primary contribution of this work lies not in demonstrating superior predictive performance over classical trust models, but in illustrating how quantum-inspired formalisms can serve as a compact and expressive representational framework

for trust dynamics. Classical approaches such as state-space models, Kalman filters, or Bayesian networks can, in principle, approximate nonlinear and context-dependent trust trajectories. However, doing so often requires explicitly enumerating additional states, interaction terms, or history-dependent transition rules. In contrast, the quantum-inspired framework naturally represents order sensitivity, contextuality, and interdependence through noncommutative updates and entanglement, without an explosion in model complexity. In this sense, the present findings should be interpreted as evidence that such a representation is compatible with real-time trust estimation, rather than as proof that it is categorically superior to classical alternatives.

The practical implications of a calibrated and responsive trust signal are significant. In applied settings such as autonomous driving, clinical decision support, emergency response, or training simulations, real-time estimates of user trust could inform adaptive interface behaviors, including adjustments to explanation depth, transparency, or autonomy allocation. Because the model exhibited minimal systematic bias, its outputs could be mapped directly to actionable thresholds (e.g., triggering additional explanations when trust drops below a predefined level) without requiring ad hoc correction. These properties align with current recommendations for designing AI systems that support appropriate reliance rather than indiscriminate trust.

From a comparative modeling perspective, the present framework should be interpreted as complementary to, rather than a replacement for, established trust-aware decision-theoretic approaches. Trust-POMDP formulations and their extensions jointly estimate trust and plan over it, demonstrating improved team performance and explicitly capturing asymmetric trust dynamics in which failures exert greater influence than successes [45], [46]. Similarly, recent IOHMM- and POMDP-based assistance models learn latent trust transition structures directly from behavioral data and validate these estimates against both performance outcomes and self-reported trust,

thereby enabling policy optimization grounded in empirically learned dynamics [47], [48]. In contrast, the current study does not seek to outperform such models nor to recover latent transition matrices or assistance policies.

Likewise, time-series approaches such as ARIMAX modeling have demonstrated clear benefits of explicitly representing temporal dependence and order effects, while dynamic structural equation models (DSEM) provide interpretable causal structure and outperform autoregressive baselines across multiple domains. Although the quantum-inspired formulation naturally encodes order sensitivity and contextual dependence through noncommutative updates and entanglement, the present evaluation does not yet include head-to-head comparisons against AR-family, DSEM, or multi-factor latent-state POMDP baselines, including joint trust-workload models that transparently represent multiple interacting latent constructs. Finally, while the broader quantum interaction literature motivates the use of quantum probability to capture contextuality and interference effects that challenge classical probabilistic models [49], future work must empirically demonstrate that these representational features yield predictive or control advantages not attainable through simpler classical formulations. As such, the contribution of this work lies in establishing the feasibility of a compact, quantum-inspired representation for real-time trust estimation in human-AI interaction, while motivating systematic comparative evaluations as a critical next step.

Several limitations of the present study warrant consideration. First, the parameter values and unitary mappings used in the model should be interpreted as engineering calibrations chosen to satisfy qualitative constraints, such as bounded updates, nonsaturation, and plausible trajectories, rather than as uniquely identified or optimal parameters. Future work could refine these mappings using formal parameter estimation or optimization techniques. Second, the sample

size was modest and demographically narrow, which limits the generalizability of the findings to broader populations. Third, the current implementation relied exclusively on textual interaction features; incorporating multimodal signals such as speech prosody, gaze behavior, or physiological measures may improve robustness and provide a more stringent test of the framework.

Additionally, while the present results demonstrate strong agreement between model estimates and self-reported trust, they are not yet contextualized against matched classical baselines or component-level ablations. The primary objective of this study was to establish feasibility within a real-time interaction loop, rather than to isolate the unique contribution of noncommutative updates through controlled manipulations of event order. Future studies should therefore compare quantum-inspired and classical models under carefully designed sequences that vary ordering, context, and cue interactions, allowing a more direct test of representational advantages.

The strong correspondence observed in short interactions may also reflect the model’s initialization strategy. By anchoring the initial quantum state to participants’ self-reported initial trust and perceived risk, the estimator begins aligned with the user’s internal state. Over relatively brief interaction sequences, this anchoring likely constrains divergence, emphasizing calibration rather than long-term drift. Longitudinal studies involving extended interactions will be necessary to evaluate how trust trajectories evolve over time and whether additional mechanisms are required to manage accumulation, decay, or recovery effects.

Finally, deploying the framework on physical quantum hardware introduces additional challenges related to noise, decoherence, and circuit depth limitations. Although the present study relies on state-vector simulation, benchmarking performance under realistic quantum operating conditions represents an important direction for future work.

In summary, this study demonstrates the feasibility of a quantum-inspired framework for real-time trust estimation in human-AI interaction. By producing calibrated and responsive trust estimates across varying task contexts, the model provides a foundation for adaptive, trust-aware AI systems. Rather than positioning quantum formalism as a replacement for classical models, the present work highlights its potential as a complementary representational approach for capturing contextual, order-dependent aspects of trust that are increasingly relevant in dynamic human-AI collaboration.

5. CONCLUSION

This work proposed and empirically evaluated an automated, quantum-inspired framework for modeling trust in real-time human-AI interactions. Unlike many prior trust models that assume linear dynamics or rely on static, self-reported measurements, the present approach leverages principles such as superposition and entanglement to represent trust as a fluid, context-dependent construct that evolves over the course of interaction.

An experimental evaluation involving 15 participants interacting with a conversational AI agent demonstrated that the proposed framework produced trust estimates that were well calibrated to participants' self-reported trust and closely tracked trust trajectories over time. Across participant-reported trust ratings, the model exhibited minimal systematic bias and strong correspondence with reported values, with comparable agreement observed across both low-risk and high-risk scenarios. These findings support the feasibility of the quantum-inspired representation for capturing dynamic trust behavior during short conversational interactions, rather than establishing definitive predictive superiority over classical trust models.

The implications of this work extend to application domains in which appropriate trust calibration is critical, including autonomous systems, clinical decision support, and collaborative

robotics. By enabling continuous monitoring of trust, the framework may support adaptive AI systems that adjust transparency, autonomy, or interaction strategies in response to evolving user trust, thereby promoting safer and more effective human-AI collaboration.

Future work will extend the framework through several dimensions. Methodologically, incorporating multimodal signals such as speech characteristics and physiological measures may provide a richer basis for trust estimation. Technically, implementing the model on quantum hardware rather than classical simulation backends will enable evaluation of practical constraints related to noise, circuit depth, and computational efficiency. More broadly, longitudinal studies involving extended interactions and diverse populations will be necessary to assess long-term trust dynamics, drift, and recovery. Overall, this study establishes a foundation for quantum-inspired trust modeling as a viable representational approach for real-time, trust-aware AI systems, complementing existing classical methods in safety-critical human-AI collaboration.

DATA AVAILABILITY

The datasets generated and analyzed during the current study are available from the corresponding author upon reasonable request. All participants provided informed consent, and data sharing is subject to institutional ethics approval and privacy considerations.

The quantum-inspired trust modeling framework was implemented in Python using the Qiskit library. The full source code, including the trust estimator, simulation pipeline, and example configuration files, is publicly available at: <https://github.com/reynashBadhwar26/Quantum-Trust-Model.git>. This repository contains sufficient documentation to reproduce the model behavior and replicate the analyses reported in this study.

769

770

REFERENCES

771

[1] Berretta, S. *et al.* Defining human-AI teaming the human-centered way: a scoping review and

772

network analysis. *Frontiers in Artificial Intelligence* **6**, (2023).

773

[2] Afroogh, S., Akbari, A., Malone, E., Kargar, M. & Alambeigi, H. Trust in AI: progress, challenges, and

774

future directions. *arXiv.org* <https://arxiv.org/abs/2403.14680> (2024).

775

[3] Lee, J. D. & See, K. A. Trust in automation: designing for appropriate reliance. *Human Factors the*

776

Journal of the Human Factors and Ergonomics Society **46**, 50–80 (2004).

777

[4] Drnec, K., Marathe, A. R., Lukos, J. R. & Metcalfe, J. S. From Trust in Automation to Decision

778

Neuroscience: Applying cognitive neuroscience methods to understand and improve interaction

779

decisions involved in human automation interaction. *Frontiers in Human Neuroscience* **10**,

780

(2016).

781

[5] Lee, J. & Moray, N. Trust, control strategies and allocation of function in human-machine systems.

782

Ergonomics **35**, 1243–1270 (1992).

783

[6] Busemeyer, J. R. & Townsend, J. T. Decision field theory: A dynamic-cognitive approach to decision

784

making in an uncertain environment. *Psychological Review* **100**, 432–459 (1993).

785

[7] Gao, J. & Lee, J. D. *EFFECT OF SHARED INFORMATION ON TRUST AND RELIANCE IN a DEMAND*

786

FORECASTING TASK. PROCEEDINGS of the HUMAN FACTORS AND ERGONOMICS SOCIETY 50th

787

ANNUAL MEETING—2006 215–216 <https://doi.org/10.1177/154193120605000302> (2006).

788

[8] Akash, K., Hu, N. W.-L., Reid, T. & Jain, N. Dynamic modeling of trust in human-machine interactions.

789

2017 American Control Conference (ACC) (2017) doi:10.23919/acc.2017.7963172.

790

[9] Xu, A. & Dudek, G. OPTIMo. *OPTIMo: Online Probabilistic Trust Inference Model for Asymmetric*

791

Human-Robot Collaborations (2015) doi:10.1145/2696454.2696492.

- [10] Chen, M., Nikolaidis, S., Soh, H., Hsu, D. & Srinivasa, S. Planning with Trust for Human-Robot Collaboration. *In Proceedings of the 2018 ACM/IEEE International Conference on Human–Robot Interaction (HRI ’18) (9 Pages)* 307–315 (2018) doi:10.1145/3171221.3171264.
- [11] Liu, J., Akash, K., Misu, T. & Wu, X. Clustering Human Trust Dynamics for Customized Real-Time Prediction. *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)* 1705–1712 (2021) doi:10.1109/itsc48978.2021.9565016.
- [12] Poornikoo, M., Gyldensten, W., Vesin, B. & Øvergård, K. I. Trust in Automation (TiA): Simulation Model, and Empirical Findings in Supervisory Control of Maritime Autonomous Surface Ships (MASS). *International Journal of Human-Computer Interaction* **41**, 7521–7548 (2024).
- [13] Rodriguez, L. R. *et al.* A review of mathematical models of human trust in automation. *Frontiers in Neuroergonomics* **4**, (2023).
- [14] Yang, X. J., Schemanske, C. & Searle, C. Toward Quantifying trust dynamics: How people adjust their trust after Moment-to-Moment interaction with automation. *Human Factors the Journal of the Human Factors and Ergonomics Society* **65**, 862–878 (2021).
- [15] Rittenberg, B. S. P., Holland, C. W., Barnhart, G. E., Gaudreau, S. M. & Neyedli, H. F. Trust with increasing and decreasing reliability. *Human Factors the Journal of the Human Factors and Ergonomics Society* **66**, 2569–2589 (2024).
- [16] Hoff, K. A. & Bashir, M. Trust in automation. *Human Factors the Journal of the Human Factors and Ergonomics Society* **57**, 407–434 (2014).
- [17] Kohn, S. C., De Visser, E. J., Wiese, E., Lee, Y.-C. & Shaw, T. H. Measurement of Trust in Automation: A Narrative review and reference guide. *Frontiers in Psychology* **12**, (2021).
- [18] Guyton, Z., Pak, R. & Rovira, E. The role of automation etiquette and task-criticality on performance, workload, automation reliance, and user confidence. *Applied Ergonomics* **125**, 104430 (2024).

815 [19] Jorge, C. C., Bouman, N. H., Jonker, C. M. & Tielman, M. L. Exploring the effect of automation failure
816 on the human's trustworthiness in human-agent teamwork. *Frontiers in Robotics and AI* **10**,
817 1143723 (2023).

818 [20] Ueno, T. *et al.* Trust in Human-AI Interaction: Scoping out models, measures, and methods. *CHI*
819 *Conference on Human Factors in Computing Systems Extended Abstracts* 1–7 (2022)
820 doi:10.1145/3491101.3519772.

821 [21] Gulati, S., McDonagh, J., Sousa, S. & Lamas, D. Trust models and theories in human–computer
822 interaction: A systematic literature review. *Computers in Human Behavior Reports* **16**, 100495
823 (2024).

824 [22] Trueblood, J. S. & Busemeyer, J. R. A quantum probability account of order effects in inference.
825 *Cognitive Science* **35**, 1518–1552 (2011).

826 [23] Wang, Z., Solloway, T., Shiffrin, R. M. & Busemeyer, J. R. Context effects produced by question
827 orders reveal quantum nature of human judgments. *Proceedings of the National Academy of*
828 *Sciences* **111**, 9431–9436 (2014).

829 [24] Pothos, E. M. & Busemeyer, J. R. Quantum cognition. *Annual Review of Psychology* **73**, 749–778
830 (2021).

831 [25] Kahr, P. K., Rooks, G., Snijders, C. & Willemsen, M. C. The Trust Recovery Journey. The Effect of
832 Timing of Errors on the Willingness to Follow AI Advice. *29th International Conference on*
833 *Intelligent User Interfaces* 609–622 (2024) doi:10.1145/3640543.3645167.

834 [26] Roeder, L. *et al.* A quantum model of trust calibration in Human–AI interactions. *Entropy* **25**, 1362
835 (2023).

836 [27] Miller, D. *et al.* Behavioral measurement of trust in automation. *Proceedings of the Human Factors*
837 *and Ergonomics Society Annual Meeting* **60**, 1849–1853 (2016).

838 [28] IBM Quantum. HGate (latest version). *IBM Quantum Documentation*
839 <https://docs.quantum.ibm.com/api/qiskit/qiskit.circuit.library.HGate>.

840 [29] IBM Quantum. RYGate (latest version). *IBM Quantum Documentation*
841 <https://docs.quantum.ibm.com/api/qiskit/qiskit.circuit.library.RYGate>.
842 [30] IBM Quantum. CRYGate (latest version). *IBM Quantum Documentation*
843 <https://docs.quantum.ibm.com/api/qiskit/qiskit.circuit.library.CRYGate>.
844 [31] Liu, B. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*
845 **5**, 1–167 (2012).
846 [32] Muir, B. M. Trust in automation: Part I. Theoretical issues in the study of trust and human
847 intervention in automated systems. *Ergonomics* **37**, 1905–1922 (1994).
848 [33] Beno, J. P. ELECTRA and GPT-4O: Cost-Effective Partners for Sentiment Analysis. *arXiv.org*
849 <https://arxiv.org/abs/2501.00062> (2024).
850 [34] Hancock, P. A. *et al.* A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction. *Human*
851 *Factors the Journal of the Human Factors and Ergonomics Society* **53**, 517–527 (2011).
852 [35] Chung, H. & Yang, X. J. *Trust Dynamics in Human-Autonomy Interaction: Uncover Associations*
853 *between Trust Dynamics and Personal Characteristics. IEEE* <https://arxiv.org/abs/2409.07406v1>
854 (2024).
855 [36] Schoeffer, J., De-Arteaga, M. & Kühl, N. Explanations, Fairness, and Appropriate Reliance in Human-
856 AI Decision-Making. *arXiv.org* 1–18 (2024) doi:10.1145/3613904.3642621.
857 [37] Bland, J. M. & Altman, Douglas G. STATISTICAL METHODS FOR ASSESSING AGREEMENT BETWEEN
858 TWO METHODS OF CLINICAL MEASUREMENT. *The Lancet* **327**, 307–310 (1986).
859 [38] Shapiro, S. S. & Wilk, M. B. An analysis of variance test for normality (complete samples). *Biometrika*
860 **52**, 591–611 (1965).
861 [39] Levene, H. Robust tests for equality of variances. In: *Olkin, I., Ed., Contributions to Probability and*
862 *Statistics, Stanford University Press, Palo Alto, 278-292.*

- [40] Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects models using lme4. *Journal of Statistical Software* **67**, (2015).
- [41] Lakens, D. Equivalence tests. *Social Psychological and Personality Science* **8**, 355–362 (2017).
- [42] Linnet, K. Evaluation of regression procedures for methods comparison studies. *Clinical Chemistry* **39**, 424–432 (1993).
- [43] Cornbleet, P. J. & Gochman, N. Incorrect least-squares regression coefficients in method-comparison analysis. *Clinical Chemistry* **25**, 432–438 (1979).
- [44] Carkeet, A. Exact parametric confidence intervals for Bland-Altman limits of agreement. *Optometry and Vision Science* **92**, e71–e80 (2015).
- [45] Chen, M., Nikolaidis, S., Soh, H., Hsu, D. & Srinivasa, S. Trust-Aware decision making for Human-Robot Collaboration: model learning and planning. *arXiv.org* <https://arxiv.org/abs/1801.04099> (2018).
- [46] Mangalindan, D. H., Rovira, E. & Srivastava, V. When to seek help: Trust-Aware Assistance-Seeking in Human-Supervised Autonomy. *ACM Transactions on Human-Robot Interaction* **15**, 1–23 (2025).
- [47] Dekarske, J., Kong, Z. & Joshi, S. Dynamic human trust modeling of autonomous agents with varying capability and strategy. *arXiv.org* <https://arxiv.org/abs/2404.19291> (2024).
- [48] Akash, K., Jain, N. & Misu, T. Toward Adaptive Trust Calibration for Level 2 Driving Automation. *2020 International Conference on Multimodal Interaction (ICMI '20)* 538–547 (2020) doi:10.1145/3382507.3418885.
- [49] Aerts, D., Czachor, M. & Sozzo, S. Quantum Interaction approach in cognition, artificial intelligence and robotics. *arXiv.org* <https://arxiv.org/abs/1104.3345> (2011).

AUTHOR CONTRIBUTION

All authors contributed to the conception and design of the study. **Reyansh Badhwar** led the development of the quantum-inspired trust model, implemented the computational framework, performed the data analysis, and drafted the manuscript. **Zehaan Walji, Parshva Dave and Dr. Junho Park** contributed to experimental design, data collection, interpretation of results, and critical revision of the manuscript. All authors reviewed and approved the final version of the manuscript.

COMPETING INTERESTS

The authors declare that they have no conflicts of interest with respect to the research, authorship, and/or publication of this article.

ACKNOWLEDGEMENT

Funding for this research was provided by a Natural Sciences and Engineering Research Council of Canada (NSERC) Bridge Fund (University of Calgary). The views and opinions expressed are those of the authors.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryInformation.docx](#)