From Evidence to Action: Auditable Prompts for Calibrating Reliance in Cooperative Driving

Yongming Zhang[*1,2], **Qi Zhang**[3], **Wenlong Zheng**[1] **& He Xu**[†1]
[1]School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing, China
[2]School of Big Data and Artificial Intelligence, Anhui Xinhua University, Hefei, China
[3]Hefei Yueming Traffic Electronic Technology Co., Ltd., Hefei, China

Abstract

Cooperative driving depends on time-critical vehicle-to-everything (V2X) alerts whose evidence is distributed, incomplete, and sometimes adversarial. A key cognitive failure is *miscalibrated reliance*: decision-makers may act on invalid alerts (overtrust) or dismiss valid ones (undertrust) under time pressure. We introduce *auditable trust prompts*, a decision-oriented interface that compresses verifiable cues into compact badges and an action recommendation (RELY/VERIFY/IGNORE). Prompts integrate cross-source corroboration (peer vehicles and roadside units (RSUs), plus on-board sensing such as on-board diagnostics (OBD) and global navigation satellite system (GNSS) cues), recency-weighted sender reliability, and provenance availability/verification status, while highlighting uncertainty triggers that warrant selective checking. In a human-in-the-loop alert assessment task manipulating evidence uncertainty, adversarial contamination, and congestion-driven availability stress, auditable prompts reduce unsafe overreliance on invalid alerts without a commensurate increase in missed reliance on valid ones. Behavioral logs show that verification increases mainly in hard regimes, suggesting a targeted verification pathway rather than blanket distrust. These findings suggest that auditability can serve as a usable cue for calibrated reliance, beyond backend accountability, in dynamic cooperative driving.

Keywords: trust calibration; reliance; selective verification; time pressure; cooperative driving; V2X alerts; provenance

Introduction

Cooperative driving increasingly relies on vehicle-to-everything (V2X) alerts to share hazards, intent, and cooperative perception. However, availability alone does not guarantee a safety benefit: within seconds, a driver (or a supervisory agent) must decide *how much to rely on a specific alert* and whether to cross-check it. This decision is cognitively demanding because evidence is distributed across multiple sources (nearby vehicles, roadside units, and local sensing), arrives asynchronously, and is observed only partially under time pressure (Kang et al., 2018; Zavvos et al., 2021). A central failure mode is *miscalibrated reliance*: people may overtrust unreliable alerts (acting when they should not) or undertrust reliable ones (forgoing the benefit of cooperation), either of which can degrade safety.

In realistic deployments, reliance calibration is hard for at least four reasons. First, sender quality is heterogeneous and non-stationary, so credibility must be tracked online rather than assumed (Din et al., 2021; Javaid et al., 2020; Yan et al., 2023). Second, congestion and disruption reshape what evidence is available *at the moment of choice*; verification attempts can therefore incur large tail-latency costs exactly when time is scarce (Trkulja et al., 2020; Yuan et al., 2022). Third, adversaries can manipulate the channel or the sensing context, producing plausible but unsafe evidence streams that are difficult to reject quickly (Zhou et al., 2024). Fourth, accountability often coexists with privacy constraints, limiting what can be revealed immediately and increasing the value of efficient, selective checks (Feng et al., 2019; J. Li et al., 2014).

Meanwhile, recent Internet of Vehicles (IoV) systems have strengthened *backend primitives* for integrity, traceability, and auditability of shared alerts, making supporting evidence verifiable in principle (Cui et al., 2021; Z. Ma et al., 2019). Yet these primitives typically surface as machine-facing scores, policies, or logs. This leaves a gap between evidence and action: users still must translate distributed, uncertain evidence into an action (*rely*, *verify*, or *ignore*) under time pressure.

To narrow this gap, we introduce *auditable trust prompts*: compact prompts that recommend one of three actions—RELY, VERIFY, or IGNORE—and expose only the minimum evidence needed to justify that recommendation. Prompts fuse (i) cross-source corroboration (V2X peers/RSUs and local sensing), (ii) sender reliability updated online with explicit recency (Din et al., 2021; Javaid et al., 2020), and (iii) provenance signals indicating whether lightweight verification artifacts are available and valid (Mizrahi et al., 2024). Prompts use short badges and uncertainty triggers to route attention toward *selective verification* without verbose explanations, when conditions warrant it, reducing the inferential gap between system evidence and human action (Sreedharan et al., 2025; Wirz et al., 2025). The design is also aligned with accountability-driven evidence pipelines in vehicular forensics and event-data recording, where compact, checkable summaries must remain usable in the moment (Strandberg et al., 2022; Yao & Atkins, 2020).

We evaluate whether auditable trust prompts improve reliance calibration in a human-in-the-loop alert assessment task that manipulates uncertainty, adversarial manipulation, and congestion. Because cooperative alerting pipelines increasingly incorporate learned components that shape perceived credibility, failures such as poisoning and backdoor behavior are relevant to the reliability of displayed evidence (Chen et al., 2022; X. Ma et al., 2025; Miao et al., 2024). Our primary

[*]Corresponding author: zyming@njupt.edu.cn
[†]Co-corresponding author: xuhe@njupt.edu.cn

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

outcomes are overtrust and undertrust rates; we additionally report verification behavior and decision latency (including tails under congestion) to test whether gains come from selective checking rather than uniformly conservative responding.

In summary, this paper makes three contributions:

- *Auditable prompting for cooperative driving.* An interface layer that maps corroboration, sender history, and provenance verification into actionable RELY/VERIFY/IGNORE recommendations.
- *Calibration-first evaluation under stress.* Overtrust and undertrust are treated as primary outcomes across uncertainty, attack, and congestion conditions.
- *Behavior pathway and cost accounting.* We test whether calibration gains arise through selective verification with practical decision-time costs, including tail latency under congestion.

Related Work

Calibrated reliance and selective verification in time-critical decisions. In cooperative driving, the user-facing problem is rarely "is the alert true?" in isolation; it is whether the alert should be acted on immediately, cross-checked, or set aside when time and attention are limited. This perspective foregrounds *reliance calibration* and *selective verification* as behavioral outcomes, especially when network disruption changes what evidence is available *at decision time* (Trkulja et al., 2020; Yuan et al., 2022). Prior work on V2X data sharing and trust governance largely targets system-level correctness or accountability, but offers limited guidance on how operators and drivers should allocate attention across competing alerts. Our work adopts a three-way action framing (RELY/VERIFY/IGNORE) and evaluates overtrust and undertrust as safety-relevant error modes.

Minimal, task-aligned explanations as decision support. In explainable and trustworthy AI, a common issue is the inferential gap between what a system can justify and what a user must decide in context. Recent arguments emphasize *minimal, task-aligned* explanations that help users act under constraints, rather than comprehensive rationales that are costly to parse (Sreedharan et al., 2025; Wirz et al., 2025). We follow this view by treating explanation as a *prompting layer*: compact cues that explicitly signal uncertainty and recency, and that route users to verification only when evidence is weak, conflicting, or stale. Relative to generic confidence displays, the goal is not to increase information volume, but to improve action selection under time pressure.

Auditability and provenance primitives for accountable V2X data. IoV systems increasingly provide backend mechanisms for integrity, traceability, and auditability of shared alerts. Ledger-backed trusted data management and consortium-style sharing improve accountability (Cui et al., 2021; Z. Ma et al., 2019), while smart contracts make sharing rules executable and inspectable (C. Li et al., 2024). To reduce verification cost, proof-shortening techniques (e.g., traffic-aware Merkle proofs) aim to make provenance checks more practical at runtime (Mizrahi et al., 2024). Automotive forensics and event-data recording further clarify what evidence should be captured to support later reconstruction and accountability (Mehrish et al., 2019; Strandberg et al., 2022; Yao & Atkins, 2020). We do not propose a new backend protocol; we focus on a cognitively grounded interface for translating these primitives into *decision-time* cues that support calibrated reliance.

Trust, privacy, and adversarial uncertainty in cooperative pipelines. Online trust tracking and lightweight reputation mechanisms support screening of unreliable participants in dynamic environments (Din et al., 2021; Javaid et al., 2020; Yan et al., 2023). At the same time, V2X deployments must balance conditional privacy with accountability, motivating authentication and privacy-preserving participation control (Feng et al., 2019; J. Li et al., 2014; J. Zhang et al., 2019). Adversaries can manipulate the channel or sensing context, producing plausible but unsafe evidence streams (Ceccato et al., 2021; Lin et al., 2024; Zhou et al., 2024), and learning-enabled components introduce poisoning/backdoor risks that further complicate perceived credibility (Chen et al., 2022; X. Ma et al., 2025; Miao et al., 2024; J. Zhang et al., 2024). Accordingly, we study prompts under manipulated uncertainty and attack, asking whether they reduce unsafe overreliance without inducing blanket distrust.

Congestion and scalability as user-facing constraints. Scalability mechanisms such as sharding and throughput-oriented ledger design help sustain system performance at scale (Ning et al., 2023; X. Zhang et al., 2022), but the cognitive bottleneck remains: what evidence stays timely and interpretable enough to guide action. Congestion control and DoS-style disruption can shift evidence availability and make verification expensive in the tail, even when backend correctness holds (Trkulja et al., 2020; Yuan et al., 2022). We therefore report decision latency (including tails) alongside verification behavior, linking system stressors to reliance calibration rather than treating performance solely as an infrastructure metric.

Method

Overview

We consider cooperative driving where an ego vehicle receives time-critical V2X alerts from multiple senders (vehicles and RSUs). For each incoming alert, a driver (or supervisory agent) must quickly choose one of three responses: RELY (act on it), VERIFY (spend time to check), or IGNORE (do not use it for action). Our method provides *auditable trust prompts*: compact, decision-oriented cues that summarize *verifiable* evidence and, when available, offer a lightweight

provenance check. The design goal is behavioral: reduce *miscalibrated reliance* (overtrust and undertrust) without forcing users into slow, always-verify strategies. Fig. 1 illustrates the evidence→prompt→action loop and the associated logging for later calibration analysis, and clarifies how experimental stressors reshape what evidence is available at decision time.

Alerts and decision-time evidence

Let $s \in \mathcal{S}$ denote a sender and $r \in \mathcal{R}$ a receiver (ego vehicle). Each alert is represented as

$$m_k = \langle s_k, e_k, \tau_k, x_k \rangle, \tag{1}$$

where e_k denotes the referenced event (e.g., hazard/intent), τ_k is the message generation time, and x_k is a compact payload.

For each received message m_k, the receiver extracts a bounded evidence vector

$$\mathbf{z}_k = \big[C_k,\ R_k,\ P_k,\ U_k\big]. \tag{2}$$

The components are chosen to match what a user can plausibly interpret under time pressure:

- *Sender history* $R_k \in [0,1]$: receiver-maintained reliability for s_k based on recent interactions (Sec. Online trust score and recency-weighted update).
- *Uncertainty triggers* $U_k \in [0,1]$: decision-time flags intended to push users toward VERIFY when evidence is stale, conflicting, or incomplete (e.g., staleness, cue mismatch, missing corroboration). U_k is not a verbose explanation; it is a compact indicator of "conditions that warrant checking."

Audit signal and provenance check

Each message is associated with a digest

$$h_k = H(m_k), \tag{3}$$

and an audit record $\ell_k = \langle h_k, s_k, \tau_k, \mathrm{meta}_k \rangle$ stored by an audit service (or consortium log). When a verification artifact π_k is available, the receiver computes a provenance score

$$P_k = \mathbb{I}[\mathsf{Verify}(h_k, \pi_k) = 1] \cdot q_k, \tag{4}$$

where $q_k \in [0,1]$ captures provenance quality (e.g., artifact freshness or validity window). Operationally, P_k supports an optional "tap-to-verify" action that is fast enough to be used selectively.

Online trust score and recency-weighted update

Each receiver maintains a sender belief $\theta_s(k) \in [0,1]$, interpreted as the receiver's current estimate that sender s is reliable for its driving-relevant alerts. We fuse the evidence vector into a bounded message-level trust score:

$$T_k = \sigma(w_C C_k + w_R R_k + w_P P_k - w_U U_k + b), \tag{5}$$

where $\sigma(\cdot)$ is the logistic function and weights are nonnegative to keep each term directionally interpretable (corroboration/history/provenance increase trust; uncertainty triggers reduce it). This form also supports clean ablations by setting selected weights to zero.

Sender belief is updated online with recency discount so that older alerts exert less influence:

$$\theta_{s_k}(k) = (1-\lambda)\,\theta_{s_k}(k-1) + \lambda \tilde{T}_k, \tag{6}$$

$$\tilde{T}_k = \exp(-\rho\,\Delta t_k)\,T_k, \qquad \Delta t_k = t_{\mathrm{now}} - \tau_k. \tag{7}$$

This update captures a practical cognitive constraint: users care most about *recent* reliability, especially when conditions (traffic, channel quality, adversary behavior) are nonstationary.

Prompt mapping and interface elements

A prompt is represented as

$$p_k = \langle a_k^\star,\ \gamma_k,\ \mathcal{B}_k,\ \mathcal{A}_k \rangle, \tag{8}$$

where $a_k^\star \in \{\textsc{rely}, \textsc{verify}, \textsc{ignore}\}$ is the recommendation, γ_k is a coarse cue (e.g., low/medium/high gauge), $\mathcal{B}_k$ is a small set of badges that expose the minimum evidence needed for that recommendation, and $\mathcal{A}_k$ is an optional provenance VERIFY handle (Sec. Audit signal and provenance check).

Action recommendation. We map trust to actions with two thresholds and a simple risk gate, reflecting that high-risk events warrant acting when trust is high, whereas ambiguous cases should route to verification:

$$a_k^\star = \begin{cases} \textsc{rely}, & T_k \ge \tau_{\mathrm{hi}} \ \wedge\ \mathrm{Risk}(e_k) \ge \eta, \\ \textsc{verify}, & \tau_{\mathrm{lo}} \le T_k < \tau_{\mathrm{hi}}, \\ \textsc{ignore}, & T_k < \tau_{\mathrm{lo}}. \end{cases} \tag{9}$$

Here $\mathrm{Risk}(e_k) \in [0,1]$ is a task-level severity proxy (e.g., derived from alert type), and η sets when immediate reliance is justified. We tune $(\tau_{\mathrm{lo}}, \tau_{\mathrm{hi}}, \eta)$ on pilot trials to minimize a weighted OTR+UTR objective, then fix them for all participants and conditions.

Badges. Badges expose evidence in a compact, glanceable form:

$$\mathcal{B}_k = \big\{(C,\ n_k/N_k),\ (H,\ \theta_{s_k}),\ (P,\ \mathrm{v/uv})\big\}, \tag{10}$$

where $(C, n_k/N_k)$ shows corroboration, (H, θ_{s_k}) summarizes sender history, and $(P, \mathrm{v/uv})$ indicates whether provenance is verified/unverified. Uncertainty triggers (encoded in U_k) act as lightweight attention cues to prefer VERIFY when needed, rather than expanding the prompt into a long explanation.

Decision logging and reference labels

Let the observed user action be $a_k \in \{\textsc{rely}, \textsc{verify}, \textsc{ignore}\}$. If verification is performed, let $v_k \in \{0,1\}$ denote the verification outcome. We log $\langle m_k, p_k, a_k, v_k \rangle$ along with context (uncertainty/manipulation/congestion condition) and response time. For evaluation, each trial has a reference validity label $y_k \in \{0,1\}$ obtained from later confirmation; y_k is not available at decision time.

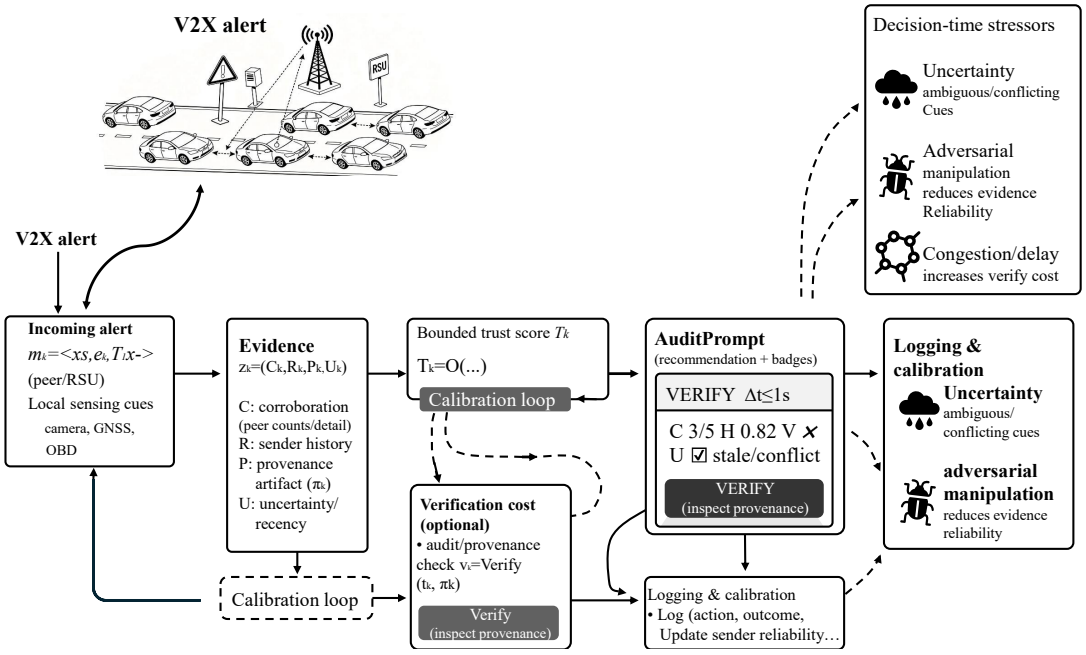

Figure 1: Auditable trust prompting loop for time-critical V2X alerts. Decision-time cues (corroboration, sender history, provenance/recency) are fused into an interpretable trust score and mapped to RELY/VERIFY/IGNORE; interactions are logged to evaluate calibration under uncertainty, attack, and congestion.

Reliance-calibration metrics and time-cost accounting

We map actions to a binary reliance indicator

$$r_k = \mathbb{I}[a_k = \text{RELY}], \tag{11}$$

and define overtrust/undertrust as

$$\begin{aligned} \text{OT}(k) &= \mathbb{I}[r_k = 1 \wedge y_k = 0]\,, \\ \text{UT}(k) &= \mathbb{I}[r_k = 0 \wedge y_k = 1]\,. \end{aligned} \tag{12}$$

Aggregated rates over a set of trials $\mathcal{K}$ are

$$\begin{aligned} \text{OTR} &= \frac{1}{|\mathcal{K}|} \sum_{k \in \mathcal{K}} \text{OT}(k), \\ \text{UTR} &= \frac{1}{|\mathcal{K}|} \sum_{k \in \mathcal{K}} \text{UT}(k). \end{aligned} \tag{13}$$

When a scalar trust score is shown, we additionally report score-calibration metrics (reliability and Brier):

$$\text{Brier} = \frac{1}{|\mathcal{K}|} \sum_{k \in \mathcal{K}} (T_k - y_k)^2. \tag{14}$$

To interpret gains under time pressure, we also record decision latency and the cost of invoking verification (communication and audit latency/bytes), reporting overhead as

$$\text{Overhead} = \text{CommCost} + \text{AuditCost}, \tag{15}$$

where AuditCost includes any verification payload and latency when $\mathcal{A}_k$ is invoked.

Online prompting loop

Algorithm 1: Online auditable prompting loop

1. Receive m_k and extract decision-time evidence $\mathbf{z}_k$ from available V2X context and local cues.
2. Compute T_k (Eq. 5) and update sender belief θ_{s_k} (Eqs. 6–7).
3. Render prompt p_k (Eq. 8) with coarse cue γ_k, badges $\mathcal{B}_k$ (Eq. 10), and optional verification handle $\mathcal{A}_k$.
4. Observe action a_k; if VERIFY, obtain v_k; log outcomes and timing for reliance-calibration analysis.

Experiments

Alert-decision task with interface cues

We test whether *auditable trust prompts* improve reliance calibration for time-critical V2X alerts under a strict decision deadline. Each trial shows one alert (hazard type, location, timestamp); participants choose RELY, VERIFY, or IGNORE.All alerts and evidence cues were scripted (no data were logged from participants' vehicles); only the interface cues varied across conditions.

Interface conditions. Alert content is identical across conditions; only decision support varies: *None* (raw alert), *Score* (raw alert + trust score T_k), *Explain* (raw alert + brief rationale, no audit action), and *AuditPrompt* (recommended action + compact badges + optional provenance VERIFY handle; Sec. Audit signal and provenance check). Table 1 summarizes

visible elements and their intended role.

Table 1: Interface conditions (same alert content; different decision support).

Condition	*What is shown*	*Decision-time intent*
None	Raw alert only.	Baseline reliance.
Score	Raw alert + trust score (T_k).	Fast triage cue.
Explain	Raw alert + short rationale (no audit action).	Minimal justification.
AuditPrompt	Recommended action + badges + optional VERIFY handle (Sec. Audit signal and provenance check); optional 3-level gauge.	Route attention to selective checks.

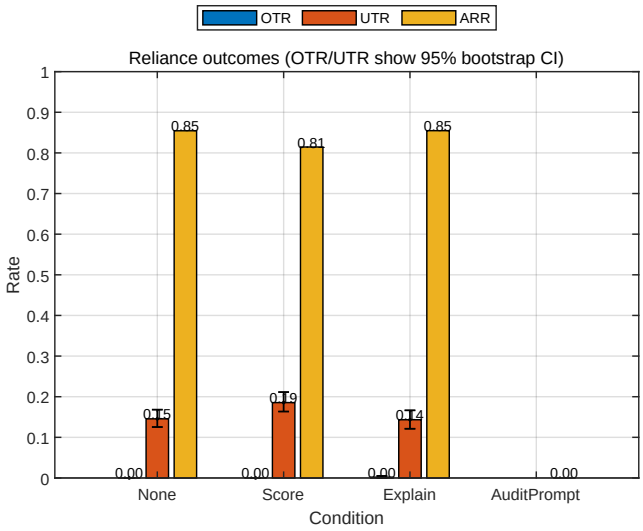

Figure 2: Overtrust (OTR) and undertrust (UTR) across interface conditions.

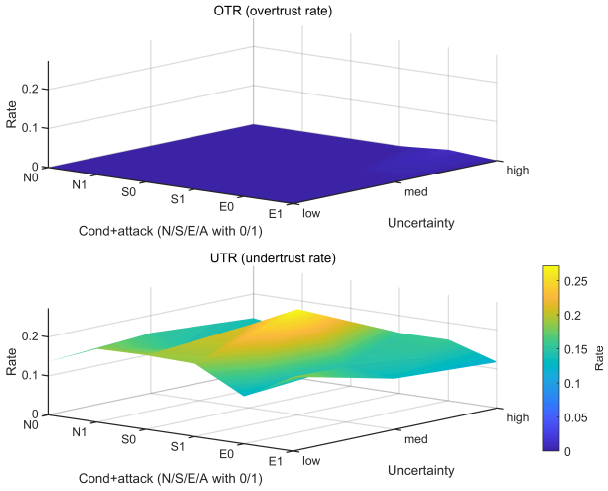

Figure 3: OTR/UTR by uncertainty and manipulation (`attack`). Color scales are shared across conditions within each metric.

Logged stressors and procedure. Trials manipulated uncertainty, `attack`, and `congestion` (affecting C_k/U_k, R_k, P_k, and verification cost). Participants (N=48) completed 6 practice trials, then four counterbalanced 30-trial blocks (one per interface; randomized within blocks) under a 1200 ms deadline (late responses logged as IGNORE); we logged condition, stressors, displayed cues, action, and response time to `trials.csv` (MATLAB R2023b; consented, anonymized).

Measures and analysis approach

Primary outcomes are overtrust, undertrust, and appropriateness (ARR); secondary outcomes include Brier score, verification rate (VR), and decision time (median and tails). We report 95% uncertainty intervals and sweep the action thresholds on the same trials to check robustness.

Reliance calibration and robustness

Overall calibration. Fig. 2 summarizes OTR/UTR by interface; improvement means lower OTR without a compensatory rise in UTR (ARR reported as a summary check). We also report an overall appropriateness rate (ARR), defined as

$$\text{ARR} = 1 - \text{OTR} - \text{UTR}. \tag{16}$$

Unlike binary accuracy, ARR keeps the two safety-relevant failure modes distinct (unsafe reliance vs. missed reliance) and does not collapse VERIFY into a "correct" decision.

Across interfaces, AuditPrompt reduces unsafe overreliance on invalid alerts while keeping missed reliance on valid alerts comparable.

Stress-localized robustness. Fig. 3 breaks errors down by uncertainty and manipulation; the high-uncertainty + `attack`=1 cells are most vulnerable to overtrust under fast decisions. Concentrated gains there suggest better allocation under ambiguity rather than uniform distrust.

The same hard regimes that concentrate overtrust errors are also where verification increases most under AuditPrompt (Fig. 6), consistent with selective checking rather than uniformly conservative responding.

Score calibration and threshold sensitivity

If a score is shown, it should be interpretable as a validity proxy and remain usable under shifting conditions. Fig. 4 compares T_k to empirical validity (closer to diagonal and lower Brier is better). Fig. 5 sweeps thresholds on T_k to check sensitivity; a better policy shifts the OTR–UTR frontier toward the bottom-left. Reliability curves use fixed binning and the sweep uses the same trial set across conditions to ensure comparability.

Selective verification and decision-time cost

We test whether prompts improve calibration by making verification *selective* in hard cases. Fig. 6 reports verification rates by uncertainty and manipulation; increases are concentrated in harder regimes.

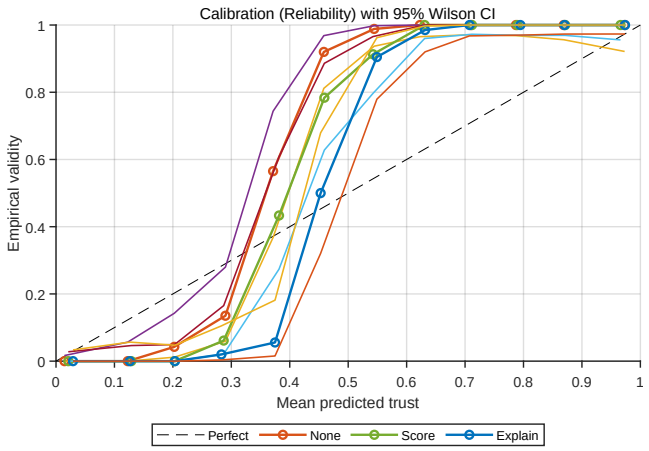

Figure 4: Reliability diagram for the underlying trust score T_k across interfaces (None/Score/Explain); only *Score* displays T_k numerically.

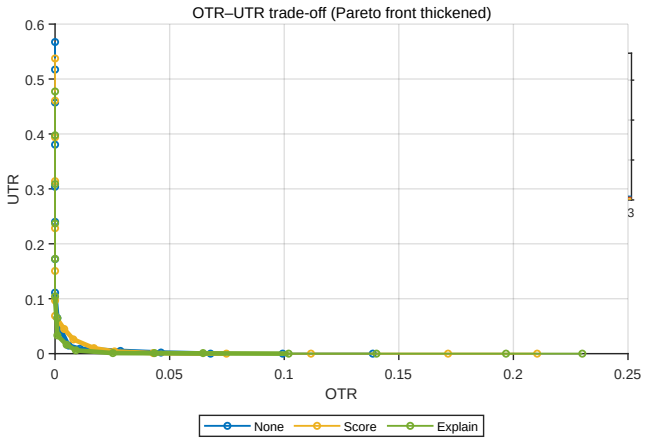

Figure 5: OTR–UTR trade-off under threshold sweeps on T_k; better policies shift toward the bottom-left.

To ensure selective checks remain practical under availability stress, Fig. 7 reports baseline decision-time latency (medians and tails) for None/Score/Explain without the tap-to-verify action, and we interpret AuditPrompt jointly with verification patterns in Fig. 6; verification increases mainly in high-uncertainty or `attack`=1 trials rather than uniformly across conditions.

Discussion and future work

Auditable prompts reduce unsafe overtrust in the most vulnerable regimes while avoiding a shift toward blanket distrust. Together, Fig. 3 and Fig. 6 indicate that verification rises mainly in harder cases under time pressure. Future work will test richer driving workload and embodied verification costs in simulator settings.

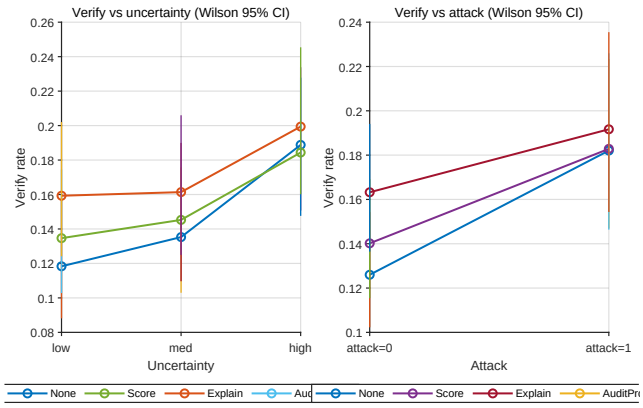

Figure 6: Verification rate by uncertainty (left) and manipulation (`attack`, right), showing increases primarily in harder regimes.

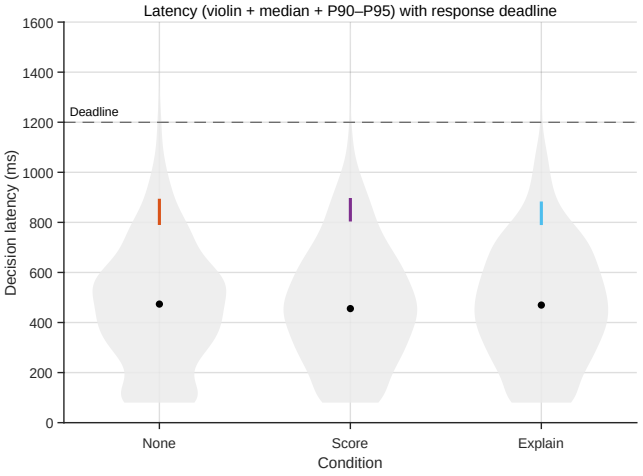

Figure 7: Decision latency distributions for baseline displays (None/Score/Explain), summarized by median and tail percentiles.

Conclusion

We introduced *auditable trust prompts* to narrow the evidence-to-action gap in time-critical V2X alerting. By converting corroboration, recency-weighted sender history, and provenance status into compact decision-time cues and an explicit RELY/VERIFY/IGNORE recommendation, prompts improve reliance calibration under uncertainty, adversarial manipulation, and congestion-driven availability stress. The observed shift toward selective verification in hard regimes suggests a practical mechanism for reducing unsafe overreliance without inducing blanket distrust. Overall, the results position auditability as a decision-support cue for calibrated reliance, beyond a post hoc accountability feature.

References

Ceccato, M., Formaggio, F., Laurenti, N., & Tomasin, S. (2021). Generalized likelihood ratio test for gnss spoof-

ing detection in devices with imu. *IEEE Transactions on Information Forensics and Security*, *16*, 3496–3509.

Chen, Y., Zhu, X., Gong, X., Yi, X., & Li, S. (2022). Data poisoning attacks in internet-of-vehicle networks: Taxonomy, state-of-the-art, and future directions. *IEEE Transactions on Industrial Informatics*, *19*(1), 20–28.

Cui, J., Ouyang, F., Ying, Z., Wei, L., & Zhong, H. (2021). Secure and efficient data sharing among vehicles based on consortium blockchain. *IEEE Transactions on Intelligent Transportation Systems*, *23*(7), 8857–8867.

Din, I. U., Bano, A., Awan, K. A., Almogren, A., Altameem, A., & Guizani, M. (2021). Lighttrust: Lightweight trust management for edge devices in industrial internet of things. *IEEE Internet of Things Journal*, *10*(4), 2776–2783.

Feng, Q., He, D., Zeadally, S., & Liang, K. (2019). Bpas: Blockchain-assisted privacy-preserving authentication system for vehicular ad hoc networks. *IEEE Transactions on Industrial Informatics*, *16*(6), 4146–4155.

Javaid, U., Aman, M. N., & Sikdar, B. (2020). A scalable protocol for driving trust management in internet of vehicles with blockchain. *IEEE Internet of Things Journal*, *7*(12), 11815–11829.

Kang, J., Yu, R., Huang, X., Wu, M., Maharjan, S., Xie, S., & Zhang, Y. (2018). Blockchain for secure and efficient data sharing in vehicular edge computing and networks. *IEEE Internet of Things Journal*, *6*(3), 4660–4670.

Li, C., Zhang, Y., Wu, J., Luo, Y., & Yu, S. (2024). Smart contract-based decentralized data sharing and content delivery for intelligent connected vehicles in edge computing. *IEEE Transactions on Intelligent Transportation Systems*, *25*(10), 14535–14545.

Li, J., Lu, H., & Guizani, M. (2014). Acpn: A novel authentication framework with conditional privacy-preservation and non-repudiation for vanets. *IEEE Transactions on Parallel and Distributed Systems*, *26*(4), 938–948.

Lin, F., Yan, H., Li, J., Liu, Z., Lu, L., Ba, Z., & Ren, K. (2024). Phade: Practical phantom spoofing attack detection for autonomous vehicles. *IEEE Transactions on Information Forensics and Security*, *19*, 4199–4214.

Ma, X., Li, X., Zhang, J., Ma, Z., Jiang, Q., Liu, X., & Ma, J. (2025). Repairing backdoor model with dynamic gradient clipping for intelligent vehicles. *IEEE Transactions on Dependable and Secure Computing*, *22*(1), 804–818.

Ma, Z., Wang, X., Jain, D. K., Khan, H., Gao, H., & Wang, Z. (2019). A blockchain-based trusted data management scheme in edge computing. *IEEE Transactions on Industrial Informatics*, *16*(3), 2013–2021.

Mehrish, A., Singh, P., Jain, P., Subramanyam, A. V., & Kankanhalli, M. (2019). Egocentric analysis of dash-cam videos for vehicle forensics. *IEEE Transactions on Circuits and Systems for Video Technology*, *30*(9), 3000–3014.

Miao, Y., Yan, X., Li, X., Xu, S., Liu, X., Li, H., & Deng, R. H. (2024). Rfed: Robustness-enhanced privacy-preserving federated learning against poisoning attack. *IEEE Transactions on Information Forensics and Security*, *19*, 5814–5827.

Mizrahi, A., Koren, N., Rottenstreich, O., & Cassuto, Y. (2024). Traffic-aware merkle trees for shortening blockchain transaction proofs [early access]. *IEEE/ACM Transactions on Networking*.

Ning, K., Liu, C., Lin, S., Geng, X., & Han, G. (2023). A blockchain sharding-based data sharing scheme for internet of vehicles. *Proceedings of the 2023 IEEE International Conferences on Internet of Things (iThings), Green Computing & Communications (GreenCom), Cyber, Physical & Social Computing (CPSCom), Smart Data (SmartData) and Congress on Cybermatics*, 7–12.

Sreedharan, S., Srivastava, S., & Kambhampati, S. (2025). Explain it as simple as possible, but no simpler–explanation via model simplification for addressing inferential gap. *Artificial Intelligence*, *340*, 104279.

Strandberg, K., Nowdehi, N., & Olovsson, T. (2022). A systematic literature review on automotive digital forensics: Challenges, technical solutions and data collection. *IEEE Transactions on Intelligent Vehicles*, *8*(2), 1350–1367.

Trkulja, N., Starobinski, D., & Berry, R. A. (2020). Denial-of-service attacks on c-v2x networks. *arXiv preprint arXiv:2010.13725*.

Wirz, C. D., Demuth, J. L., Bostrom, A., Cains, M. G., Ebert-Uphoff, I., Gagne II, D. J., Schumacher, A., McGovern, A., & Madlambayan, D. (2025). (re) conceptualizing trustworthy ai: A foundation for change. *Artificial Intelligence*, 104309.

Yan, K., Zeng, P., Wang, K., Ma, W., Zhao, G., & Ma, Y. (2023). Reputation consensus-based scheme for information sharing in internet of vehicles. *IEEE Transactions on Vehicular Technology*, *72*(10), 13631–13636.

Yao, Y., & Atkins, E. (2020). The smart black box: A value-driven high-bandwidth automotive event data recorder. *IEEE Transactions on Intelligent Transportation Systems*, *22*(3), 1484–1496.

Yuan, M., Xu, Y., Zhang, C., Tan, Y., Wang, Y., Ren, J., & Zhang, Y. (2022). Trucon: Blockchain-based trusted data sharing with congestion control in internet of vehicles. *IEEE Transactions on Intelligent Transportation Systems*, *24*(3), 3489–3500.

Zavvos, E., Gerding, E. H., Yazdanpanah, V., Maple, C., Stein, S., et al. (2021). Privacy and trust in the internet of vehicles. *IEEE Transactions on Intelligent Transportation Systems*, *23*(8), 10126–10141.

Zhang, J., Zhu, C., Sun, X., Ge, C., Chen, B., Susilo, W., & Yu, S. (2024). Flpurifier: Backdoor defense in federated learning via decoupled contrastive training. *IEEE Transactions on Information Forensics and Security*, *19*, 4752–4766.

Zhang, J., Cui, J., Zhong, H., Chen, Z., & Liu, L. (2019). Pacrt: Chinese remainder theorem based conditional privacy-preserving authentication scheme in vehicular ad-hoc networks. *IEEE Transactions on Dependable and Secure Computing*, *18*(2), 722–735.

Zhang, X., Xia, W., Cui, Q., Tao, X., & Liu, R. P. (2022). Efficient and trusted data sharing in a sharding-enabled vehicular blockchain. *IEEE Network*, *37*(2), 230–237.

Zhou, M., Zhou, W., Huang, J., Yang, J., Du, M., & Li, Q. (2024). Stealthy and effective physical adversarial attacks in autonomous driving. *IEEE Transactions on Information Forensics and Security*, *19*, 6795–6809.