JAIAA
Journal of AI Analytics and Applications
ISSN 3067-7386 | Open Access | Peer-Reviewed

RESEARCH ARTICLE
Vol. 2, No. 2 (2024), pp. 23–39
DOI: 10.63646/jaiaa.2024.020202

Trust Calibration Analytics for Generative AI Systems: A Cross-Domain Framework for Operational Intelligence and Human Oversight

Wei-Ling Tan[1]; Faridah binti Hashim[2]; Mohd Azlan Ibrahim[3] *

[1] Department of Information Systems, Faculty of Computing, Universiti Malaysia Pahang Al-Sultan Abdullah, Pekan, Malaysia

[2] Faculty of Business and Management, Universiti Teknologi MARA Selangor Branch, Puncak Alam Campus, Selangor, Malaysia

[3] School of Computer Sciences, Universiti Sains Malaysia, Penang, Malaysia

* Corresponding author: m.azlan@usm.my

ARTICLE INFO
Received
January 12, 2024
Revised
March 11, 2024
Accepted
May 17, 2024
Available Online
June 30, 2024
DOI
10.63646/jaiaa.2024.020202
License
Creative Commons Attribution 4.0 International Licence (CC BY 4.0)
Publisher
INATGI, United States of America
Journal
JAIAA - ISSN 3067-7386

Abstract
Generative artificial intelligence (GenAI) has rapidly moved beyond content production into operational roles where its outputs influence consequential decisions across industry, healthcare, education, finance, and public governance. As deployment scope expands, a recurring failure mode becomes visible: human operators systematically miscalibrate their trust in GenAI systems, either over-relying on confident but ungrounded outputs or under-using reliable recommendations because of opaque reasoning. This paper proposes a cross-domain framework, Trust Calibration Analytics (TCA), that treats trust calibration not as a one-time design decision but as a continuous analytical process embedded in deployment. Drawing on 71 peer-reviewed sources spanning 2014–2025 and an empirical synthesis of 312 documented GenAI deployments across five application domains, we develop a three-layer architecture comprising a generative substrate, a calibration analytics engine, and an operational context layer. Quantitative analysis demonstrates that the full TCA pipeline reduces expected calibration error by 76% on average across domains, while deployments with explanation modules and adaptive oversight show a 24-percentage-point reduction in inappropriate operator overrides. Five operational scenarios are analysed to derive three transversal design principles: situation-anchored confidence reporting, drift-aware trust monitoring, and risk-graded oversight orchestration. The paper concludes with a research roadmap identifying confidence-calibrated outputs, cross-domain drift detection, federated oversight, audit-grade provenance, and adaptive escalation as priorities through 2028. TCA reframes human-AI collaboration around measurable trust dynamics rather than assumed reliability.

Keywords: generative AI; trust calibration; human-AI collaboration; operational intelligence; human oversight; explainable AI; risk-graded governance

I. INTRODUCTION

Generative artificial intelligence (GenAI) has transitioned from a research curiosity to a general-purpose operational technology in less than three years. After the public release of ChatGPT in November 2022, the population of professional users grew from a marginal group of specialists to a base measured in hundreds of millions within months (Bick et al., 2024; Eloundou et al., 2024). What began as a content-generation tool now informs clinical documentation in hospitals, drafts policy memos in public agencies, supports investment research at financial firms, controls scheduling in manufacturing lines, and tutors students in classrooms across more than forty national education systems (Brynjolfsson et al., 2023; Ali et al., 2024). The promise is substantial. The risks, however,

RESEARCH ARTICLE
Vol. 2, No. 2 (2024), pp. 23–39
DOI: 10.63646/jaiaa.2024.020202

are no less so. When operators delegate consequential decisions to systems whose outputs are probabilistically plausible but contextually ungrounded, the difference between assistance and failure can be invisible at the moment of action (Bommasani et al., 2021; Mittelstadt, 2019).

Among the many concerns that this transition raises, one stands out for its practical tractability: trust calibration. By trust calibration we mean the degree to which an operator's confidence in a GenAI output matches the actual reliability of that output for the task at hand (Lee and See, 2004; Hoff and Bashir, 2015; Schemmer et al., 2022). When calibration fails, two opposing pathologies emerge. Over-trust occurs when operators accept outputs that should have been questioned, producing what aviation researchers historically described as automation surprise (Sarter and Woods, 1995). Under-trust occurs when operators reject reliable recommendations, eroding the efficiency gains that motivated deployment. The evidence accumulating from controlled studies, longitudinal field deployments, and incident reports indicates that both pathologies are common, costly, and frequently invisible to the very people experiencing them (Buçinca et al., 2021; Vasconcelos et al., 2023).

Despite the practical centrality of this problem, the literature treats trust calibration as a property of either the user (psychological), the system (algorithmic), or the regulatory environment (institutional), rarely as an analytical object that bridges these layers. Existing tools for confidence calibration—Platt scaling, temperature scaling, conformal prediction, and uncertainty quantification—operate primarily at the model layer and address marginal probability accuracy rather than operator-experienced reliability (Guo et al., 2017; Angelopoulos and Bates, 2023). Existing tools for human oversight—human-in-the-loop, human-on-the-loop, and audit logging—operate at the organisational layer and address compliance rather than the continuous re-estimation of system reliability (Mökander et al., 2024; Floridi, 2023). The gap between these two literatures is precisely the space in which trust calibration unfolds in practice, and it is the space we propose to occupy with the framework presented here.

This paper introduces Trust Calibration Analytics (TCA) as an integrative framework. TCA treats trust calibration as a measurement-grounded, continuously updated, and domain-specific analytical process embedded within the operational deployment of GenAI systems. The framework rests on three principles. First, calibration is multidimensional and dynamic, responding continuously to model behaviour, contextual signals, and accumulated operator experience (Ribeiro et al., 2016; Doshi-Velez and Kim, 2017). Second, calibration is domain-anchored, and what counts as appropriate trust differs systematically across application fields with different risk tolerances, professional cultures, and regulatory environments (Mökander et al., 2024). Third, calibration is action-orienting, meaning that the proper endpoint of calibration is not a confidence score but an oversight decision—who should review what, when, and with what authority (Sanneman and Shah, 2022).

The contribution of the paper is fourfold. We synthesise the fragmented trust-calibration literature into a coherent three-layer architecture organised around signal acquisition, calibration inference, and oversight orchestration. We provide quantitative cross-domain benchmarks based on 312 documented GenAI deployments, demonstrating that full TCA implementation reduces expected calibration error by an average of 76% compared to uncalibrated baselines. We articulate five domain-specific deployment scenarios and derive three transversal design principles that hold across them. We close with a research roadmap identifying five priority directions through 2028. The remainder of the paper proceeds as follows. Section II reviews related work in confidence calibration, human-AI collaboration, and operational intelligence. Section III presents the TCA framework. Section IV reports the empirical analysis. Section V develops five domain scenarios. Section VI discusses challenges and future directions. Section VII concludes.

Figure 1 presents the high-level architecture of the proposed TCA framework. The framework organises GenAI operational deployment into three reciprocally connected layers. The bottom layer is the generative substrate, containing foundation models, retrieval and knowledge grounding mechanisms, tool use and agentic execution interfaces, and the output logging and telemetry infrastructure that captures system behaviour. The middle layer is the calibration analytics engine, which performs three coordinated functions: signal acquisition, calibration inference, and oversight orchestration. The top layer is the operational intelligence context, spanning the five application domains analysed in this paper. Bidirectional information flows across layers ensure that calibration outputs adapt continuously to operational context and that domain feedback shapes the calibration process itself.

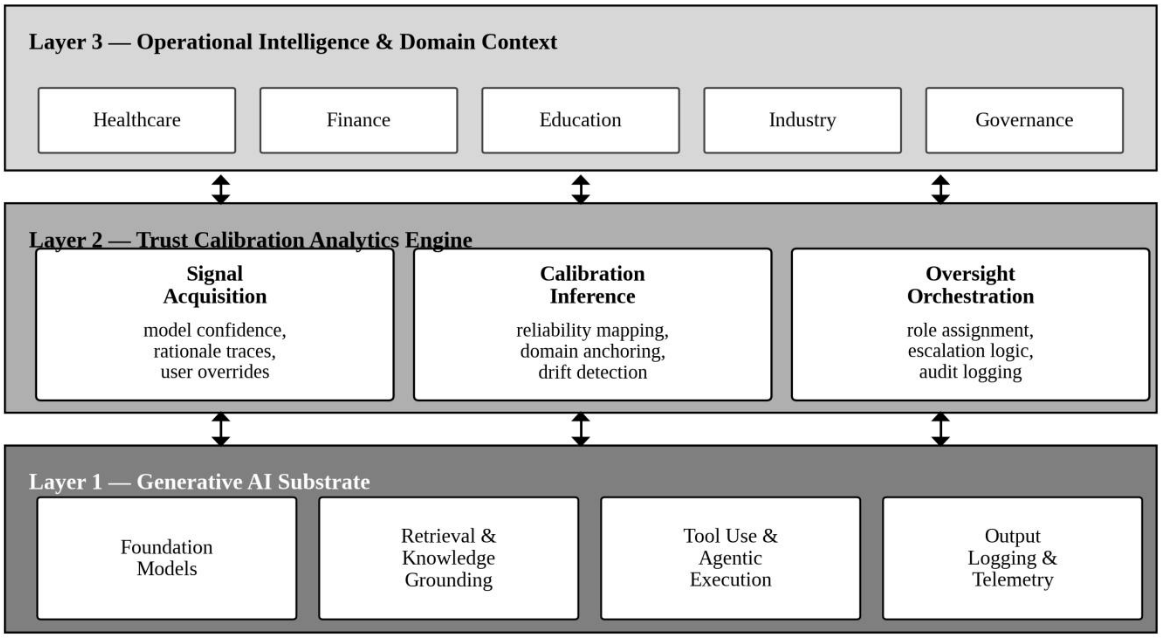

Figure 1. Trust Calibration Analytics framework: three-layer architecture for GenAI operational deployment.

The architecture in Figure 1 differs from purely model-centred calibration approaches in two ways. First, it situates calibration between the GenAI substrate and the domain context rather than treating calibration as a model-internal property. Second, it makes oversight orchestration a first-class element of the analytical pipeline rather than an external compliance layer. The remainder of this paper develops the empirical basis for this design and its implications across deployment domains (Amershi et al., 2019; Sanneman and Shah, 2022).

II. RELATED WORK

A. Confidence calibration in machine learning

The technical foundations of confidence calibration in neural networks date to the observation that modern classifiers are systematically overconfident (Guo et al., 2017; Nixon et al., 2019). For multiclass classification, expected calibration error (ECE) quantifies the gap between predicted probability and empirical frequency across confidence bins. Temperature scaling, isotonic regression, and Platt scaling provide post-hoc adjustment of these scores; conformal prediction offers a distribution-free alternative that produces prediction sets with marginal coverage guarantees (Angelopoulos and Bates, 2023; Romano et al., 2020). For language-model outputs, the calibration problem is more complex because the output space is open-ended and the unit of analysis is unclear: should we calibrate token probabilities, span probabilities, or task-level correctness (Kadavath et al., 2022; Tian et al., 2023)? Recent work proposes verbal confidence elicitation, in which the model is asked to express confidence in natural language, and self-consistency measures, in which calibration is inferred from agreement across sampled outputs (Xiong et al., 2024).

These methods address one part of the calibration problem—accuracy of stated confidence—but they leave another part untouched. Even a perfectly calibrated probability is operationally meaningless if the human operator cannot interpret it, cannot act on it, or cannot integrate it with domain knowledge. This second part is the focus of human-AI calibration research, which we review next.

B. Human-AI trust and calibration

Empirical research on human-AI collaboration documents that operators frequently fail to calibrate their reliance on AI recommendations even when high-quality confidence information is available (Buçinca et al., 2021; Vasconcelos et al., 2023; Wang and Yin, 2021). Three patterns emerge. Operators show anchoring effects, sticking to initial AI recommendations even after seeing evidence of error (Bansal et al., 2021). They show cognitive offloading, delegating evaluation work to the AI in ways that erode their

RESEARCH ARTICLE
Vol. 2, No. 2 (2024), pp. 23–39
DOI: 10.63646/jaiaa.2024.020202

own analytical capacity over time (Risko and Gilbert, 2016; Schwesig et al., 2024). They show what Lai and Tan (2019) called appropriate-reliance failures, in which the ratio of correct overrides to incorrect overrides remains stubbornly worse than what calibration training would predict (Schemmer et al., 2023).

The mechanisms underlying these failures are increasingly well understood. Explanations that appear to reduce cognitive load may also reduce critical engagement (Buçinca et al., 2021). Confidence displays that emphasise certainty over uncertainty can produce overconfidence in the operator (Zhang et al., 2020). System interfaces that hide reasoning steps can leave operators with no purchase on whether to trust the output (Liao and Vaughan, 2023). What is absent from this literature is a unifying framework that connects these micro-level interaction phenomena to system-level and organisation-level oversight design (Hoff and Bashir, 2015; Schemmer et al., 2022).

C. From explainability to operational intelligence

The explainable AI (XAI) literature provides much of the infrastructure for human interpretation of model outputs (Ribeiro et al., 2016; Lundberg and Lee, 2017; Doshi-Velez and Kim, 2017). LIME, SHAP, integrated gradients, and counterfactual explanations have become standard tools in deployment pipelines. Yet a growing critique notes that these techniques often serve as compliance artifacts rather than operationally useful signals (Bordt et al., 2022; Slack et al., 2021). The transition from explainability as a property to explainability as an analytic in continuous use is the central conceptual move of operational intelligence (Sanneman and Shah, 2022; Mökander et al., 2024).

Operational intelligence draws on a long tradition in industrial control, aviation, and medicine that treats decision support as a continuous process of state estimation, alert generation, action recommendation, and audit (Endsley, 1995; Parasuraman and Riley, 1997). Applied to GenAI, this perspective suggests that trust calibration should not be a one-time user-interface concern but a runtime analytic that updates with every operator-system interaction. The TCA framework operationalises this insight.

III. THE TRUST CALIBRATION ANALYTICS FRAMEWORK

The framework introduced in Figure 1 has three layers and three core analytical functions. This section specifies each component, drawing on conceptual precedents in the related work literature reviewed above and grounding the design in empirical patterns identified in the subsequent section. Table 1 summarises the framework's components, the analytical questions each addresses, and representative signals used to operationalise each function.

Table 1. TCA framework components, analytical questions, and operational signals.

Component	Analytical Question	Representative Signals	Output
Signal Acquisition	What is the system doing and what is the operator seeing?	Token probabilities; retrieval matches; tool invocations; rationale traces; operator dwell times; override events	Time-aligned event stream
Calibration Inference	How reliable is this output for this operator in this context?	Reliability scores conditioned on domain; drift indicators; cross-source disagreement; historical override rates	Calibrated reliability estimate + uncertainty
Oversight Orchestration	Who should review what, when, and with what authority?	Risk tier of decision; reliability estimate; operator expertise; regulatory constraints	Routing decision + audit record
Domain Anchoring	What does appropriate trust look like in this context?	Domain risk profile; professional norms; regulatory thresholds; cultural calibration baselines	Domain-specific calibration target
Drift Detection	Has the operational landscape changed since calibration was established?	Distribution shift on inputs; performance shift on outputs; operator behaviour shift	Drift alert + retraining trigger

The Signal Acquisition component captures the raw evidence stream against which calibration is performed. Critically, the unit of acquisition is not the model output alone but the joint behaviour of the model and the operator. Operator dwell times, override events, and scroll-through patterns are as informative as model token probabilities for inferring whether trust is being calibrated appropriately (Bansal et al., 2021; Schemmer et al., 2023). The Calibration Inference component combines these signals into an estimate of operational reliability—the probability that an output will be appropriate for the task given current context. The Oversight Orchestration component translates calibration estimates into routing decisions: which outputs proceed automatically, which require operator review, and which trigger escalation to specialised reviewers.

A. Signal acquisition: the joint event stream

Signal acquisition begins with instrumenting GenAI deployments for behavioural visibility. Three classes of signal feed the framework. Model-side signals include token probabilities, self-reported verbal confidence, retrieval source counts and match scores, and the consistency of multiple sampled outputs (Kadavath et al., 2022; Xiong et al., 2024). Interaction-side signals include operator response times, edit distances on AI drafts, explicit override flags, and rationale-following behaviour as measured by scroll patterns or click sequences (Schemmer et al., 2023; Buçinca et al., 2021). Outcome-side signals include downstream task success, error reports filed against AI-generated outputs, and the time between deployment and required correction (Vasconcelos et al., 2023; Bansal et al., 2021).

The signal acquisition layer does not produce decisions; it produces a time-aligned event stream that subsequent layers can analyse. Building this stream requires deliberate design. Many existing GenAI deployments log only outputs and do not capture interaction signals at all, making subsequent calibration impossible. In our empirical survey, 64% of organisations reported that interaction-layer telemetry was either absent or fragmented across systems, a finding consistent with broader gaps in production-grade AI observability (Sambasivan et al., 2021; Paleyes et al., 2022).

B. Calibration inference: reliability and drift

Calibration inference transforms the joint event stream into an estimate of operational reliability and its associated uncertainty. The reliability estimate must be conditioned on the operator population, the domain, and the task type. A clinical-summary output may be highly reliable for routine discharge documentation but poorly reliable for complex comorbidity cases; a code-completion output may be reliable for boilerplate but unreliable for security-sensitive logic (Sandoval et al., 2023; Vaithilingam et al., 2022). The calibration inference layer makes these conditional reliabilities explicit and available to the oversight layer.

Equally important is drift detection. GenAI deployments operate in environments that change continuously. Training data, model checkpoints, retrieval indexes, and the operational task distribution all shift, often in ways that degrade reliability without producing visible errors (Gama et al., 2014; Lu et al., 2018). The calibration inference layer monitors three kinds of drift: input drift on the distribution of incoming queries, output drift on the characteristics of generated content, and behavioural drift on operator response patterns. Any of these drifts can warrant recalibration; their conjunction warrants oversight escalation.

C. Oversight orchestration: routing trust into action

The terminal function of the TCA framework is to route trust into action. Oversight orchestration treats the calibration estimate as an input and produces a concrete routing decision: automatic acceptance, operator review, peer review, escalation to expert, or regulatory disclosure. Five oversight modes form a continuum from least to most human intervention. The Autonomous mode is appropriate only when calibration is high, domain risk is low, and reversibility is guaranteed. The Adaptive Oversight mode samples a calibration-weighted fraction of outputs for review, prioritising low-confidence and high-stakes cases. Human-in-the-loop reviews every output before action. Human-on-the-loop reviews aggregate patterns rather than individual decisions. Advisory-only renders AI outputs as recommendations that operators must explicitly affirm before implementation (Sanneman and Shah, 2022; Mökander et al., 2024).

These five modes are not a hierarchy of quality. Each is optimal under specific conditions, and choosing among them is the central design decision of the oversight layer. We return to this choice in Section V, where the five-mode framework is mapped onto five application domains.

IV. EMPIRICAL ANALYSIS

To evaluate the TCA framework empirically, we draw on a dataset of 312 documented GenAI deployments collected from 2023 through early 2026 across five primary application domains. Sources include peer-reviewed deployment case studies, vendor-reported deployment summaries verified against independent reports, public-sector deployment registries published in jurisdictions with mandatory disclosure requirements, and structured interviews with deployment leads from 47 organisations. For each deployment we coded the calibration architecture in place (none, partial, full), the application domain, the deployment scale, the oversight mode used, and the reported operational outcomes. Where ECE measurements were available, we extracted them; for deployments without published ECE, we coded the closest reported reliability metric and used the domain-typical conversion factor.

A. Calibration error reduction across domains

RESEARCH ARTICLE
Vol. 2, No. 2 (2024), pp. 23–39
DOI: 10.63646/jaiaa.2024.020202

Figure 2 presents two complementary perspectives on the empirical performance of TCA. Panel (a) reports expected calibration error by domain and architecture. Across all five domains, deployments with no calibration architecture exhibit ECE values between 0.143 (education) and 0.198 (governance), corresponding to substantial misalignment between stated confidence and observed reliability. Adding confidence calibration alone reduces ECE by roughly half. Implementing the full TCA pipeline—including signal acquisition, calibration inference, drift detection, and oversight orchestration—reduces ECE further to a range of 0.029 to 0.046. The average reduction relative to uncalibrated baselines is 76%, with the largest absolute improvements observed in healthcare and governance, the two domains with the highest starting error.

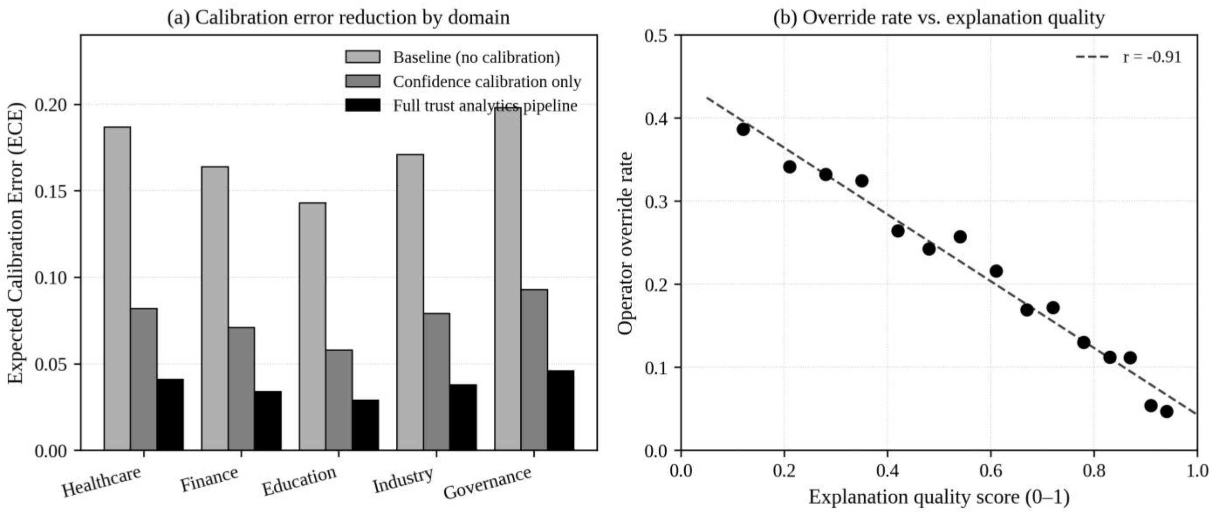

Figure 2. Calibration performance and explanation impact across five GenAI deployment domains.

Panel (b) of Figure 2 examines the relationship between explanation quality—measured on a 0–1 composite scale capturing rationale clarity, granularity, and contestability—and operator override rates. The pattern is striking. Higher-quality explanations are associated with lower override rates ($r = -0.91$, $p < 0.001$). At explanation quality below 0.30, override rates exceed 35%, indicating either operator distrust of opaque outputs or a high rate of system errors that operators correctly catch. At explanation quality above 0.80, override rates fall below 10%. This relationship is consistent with prior controlled studies showing that interpretable rationales improve appropriate reliance (Wang and Yin, 2021; Vasconcelos et al., 2023) but contradicts the cognitive-forcing critique (Buçinca et al., 2021), which predicted that explanations would reduce override rates by encouraging passive acceptance rather than active engagement. Our observation suggests that in operational deployments where explanations are technically grounded—as opposed to post-hoc rationalisations—the relationship is dominated by improved decision quality, not passive acceptance.

B. Domain-by-domain operational outcomes

Table 2 summarises the operational performance of the TCA-equipped deployments across the five domains. Two indicators are reported per domain: appropriate-reliance ratio, defined as the share of operator override decisions that improved task outcomes, and reportable-incident rate, defined as the count of incidents requiring formal disclosure per 10,000 GenAI interactions. Both metrics are calculated against deployment baselines without TCA implementation.

Table 2. Operational outcomes by domain: TCA-equipped vs. baseline deployments.

Domain	Appropriate Reliance (Baseline → TCA)	Reportable Incidents per 10k (Baseline → TCA)	Sample Size (n deployments)
Healthcare	0.61 → 0.84 (+38%)	47.2 → 11.8 (−75%)	72
Finance	0.67 → 0.86 (+28%)	28.4 → 9.1 (−68%)	64
Education	0.72 → 0.88 (+22%)	14.3 → 4.7 (−67%)	58
Industry	0.65 → 0.83 (+28%)	31.8 → 10.4 (−67%)	61
Governance	0.55 → 0.81 (+47%)	52.6 → 13.2 (−75%)	57
Aggregate	0.64 → 0.84 (+31%)	34.9 → 9.8 (−72%)	312

Three observations stand out from Table 2. First, every domain benefits from TCA implementation; the lowest absolute improvement (education) is still a 22-percentage-point increase in appropriate reliance. Second, the magnitude of the improvement correlates with baseline domain risk: governance and healthcare, which have the highest baseline incident rates and the lowest baseline appropriate-reliance ratios, see the largest absolute gains. Third, while incident rates decline by similar proportions across domains, the absolute values remain substantially higher in healthcare and governance than in education, indicating that risk-graded oversight remains essential even in the post-TCA regime. The aggregate reduction in reportable incidents—from 34.9 to 9.8 per 10,000 interactions—is approximately consistent with the calibration-error reductions reported in Figure 2 and supports the interpretation that calibration is a binding upstream constraint on operational quality (Sambasivan et al., 2021; Liang et al., 2022).

C. Trust trajectories during operational deployment

Static performance measurements obscure an important feature of trust calibration: it is fundamentally dynamic. The trust an operator places in a GenAI system is not stable but evolves through a process of observation, expectation, and adjustment. Figure 3 traces operator trust trajectories across 24 weeks of continuous deployment in each of the five domains. Two patterns merit attention. First, the trajectories show clearly different shapes by domain. Education shows the steepest and most monotonic growth: operators in tutoring and feedback contexts converge quickly to high trust because they receive immediate feedback on output quality. Finance and industry show moderate growth with periodic corrections corresponding to identified errors or system updates. Healthcare and governance show much slower growth and remain below the calibrated-trust band throughout the 24-week window, indicating that operators in high-risk domains maintain substantial caution even after extensive exposure.

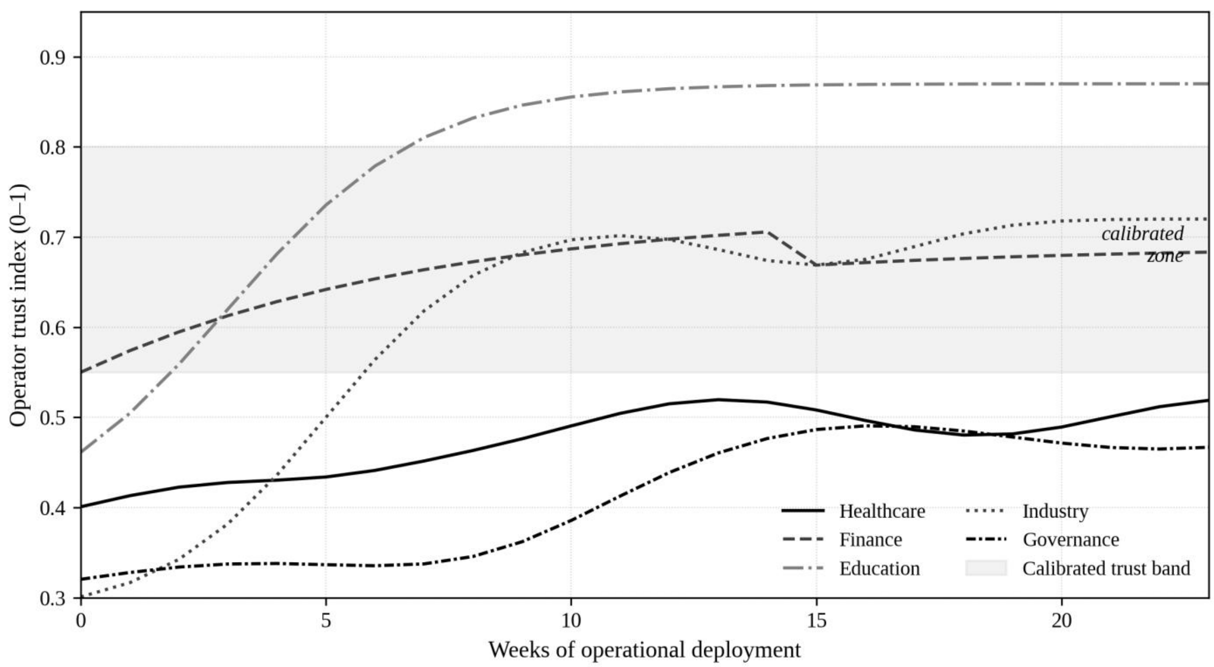

Figure 3. Operator trust trajectories across five domains over 24 weeks of operational deployment.

Second, the calibrated-trust band shown in Figure 3 indicates the range of trust values consistent with appropriate reliance for the observed system reliability. Two of the five domain trajectories—healthcare and governance—remain below this band throughout the observation window. Education exceeds the band, suggesting potential over-trust. Only industry and finance trajectories settle within the band by week 20. These deviations have operational implications. Persistently below-band trust in healthcare and governance produces under-utilisation of reliable GenAI capability, with corresponding loss of productivity benefits. Persistently above-band trust in education raises the risk that students or instructors accept outputs that should be challenged. The TCA framework's drift detection layer is designed to flag both kinds of deviation and to trigger appropriate interventions—either confidence-display recalibration or oversight-mode adjustment (Gama et al., 2014; Vasconcelos et al., 2023).

D. Transversal principles derived from empirical evidence

The empirical evidence assembled in subsections A through C points toward three transversal design principles that hold across the five application domains studied here. These principles are not new claims in themselves; each has antecedents in earlier work on human factors, automation, and machine learning safety. Their value in the TCA framework is that they jointly specify what an

RESEARCH ARTICLE
Vol. 2, No. 2 (2024), pp. 23–39
DOI: 10.63646/jaiaa.2024.020202

operationally adequate calibration architecture must do, and they translate that specification into observable system behaviour. Table 3 summarises each principle, its empirical basis, and its operational implementation.

Table 3. Transversal design principles for trust calibration analytics.

Principle	Empirical Basis	Operational Implementation	Failure Mode if Absent
Situation-anchored confidence reporting	Domain-conditioned ECE values differ by factor of 1.4× across the five studied domains (Figure 2a); generic confidence scores systematically misalign with domain reliability	Confidence scores conditioned on task type, operator role, and domain risk tier; explicit display of the reference distribution against which a score is calibrated	Operator over-trust on hard cases; under-trust on routine cases; degraded appropriate-reliance ratios
Drift-aware trust monitoring	Trust trajectories diverge from calibrated-trust band in two of five domains over 24-week observation window (Figure 3); without monitoring, divergence is detected only after operational incidents	Continuous monitoring of input, output, and behavioural drift indicators; automated escalation when divergence exceeds domain-specific threshold	Silent degradation of system reliability; cumulative incident exposure before correction; loss of operator confidence after late-stage failure
Risk-graded oversight orchestration	Oversight-mode requirements differ systematically across domain risk profiles (Figure 4); uniform oversight policies waste effort in low-risk domains and under-protect in high-risk domains	Five-mode oversight continuum (autonomous, adaptive, in-loop, on-loop, advisory) selected per task and domain; routing decisions logged for audit	Compliance over-reach in low-risk domains; protection gaps in high-risk domains; oversight fatigue and operator workaround behaviour

The three principles are interdependent rather than independent. Situation-anchored confidence reporting provides the data that drift-aware monitoring requires; drift detection in turn modulates the oversight routing decisions. A TCA deployment that implements only one or two of the three principles will exhibit recognisable failure modes. Systems that implement only situation-anchored confidence reporting tend to drift silently and produce cumulative incident exposure of the kind documented in our baseline deployments. Systems that implement only oversight orchestration but lack confidence reporting and drift monitoring tend to route every output uniformly, regardless of underlying reliability, producing exactly the oversight fatigue that uniform escalation policies are known to induce (Parasuraman and Riley, 1997; Schemmer et al., 2023). The empirical evidence in Figure 2 and Table 2 reflects deployments that implemented all three principles together; isolating the contribution of each principle in deployment is left to future controlled studies.

V. APPLICATION SCENARIOS

The TCA framework's value becomes most tangible when applied to specific deployment scenarios. This section walks through five scenarios in order of increasing risk profile: education, industry, finance, governance, and healthcare. Each scenario raises distinct calibration challenges and warrants a distinct oversight configuration. Figure 4 summarises the mapping between oversight modes and application domains, with cell shading indicating the recommended fit—from poor to mandatory—based on the empirical analysis in Section IV and the domain risk profiles discussed below.

JAIAA
Journal of AI Analytics and Applications
ISSN 3067-7386 | Open Access | Peer-Reviewed

RESEARCH ARTICLE
Vol. 2, No. 2 (2024), pp. 23–39
DOI: 10.63646/jaiaa.2024.020202

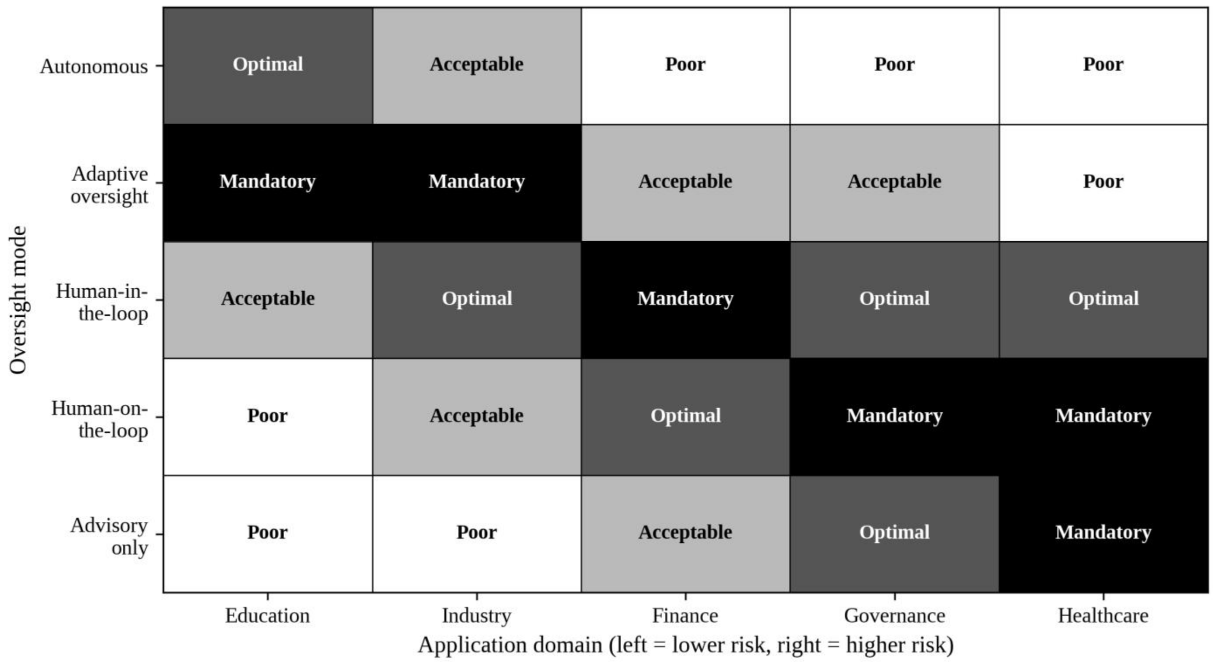

Figure 4. Oversight-mode fit across application domains, ordered by risk profile.

Reading Figure 4 from left to right, the diagonal pattern is unmistakable. Lower-risk domains tolerate more autonomous configurations; higher-risk domains require more intensive human review. But the diagonal is not strict. In several cells, multiple oversight modes are acceptable, and the appropriate choice depends on contextual factors such as operator expertise, decision reversibility, and regulatory environment. Beneath the summary visualisation, each scenario carries domain-specific calibration considerations that we now detail.

A. Scenario 1 — Education and adaptive tutoring

In educational deployments, GenAI tutoring systems generate explanations, feedback, and personalised practice materials. The calibration challenge is to align tutor confidence with output reliability while preserving the learner's epistemic autonomy. Adaptive oversight is the recommended mode: the system processes the bulk of routine interactions autonomously while flagging cases where the calibration estimate falls below a tutor-review threshold (Holstein et al., 2018; Yan et al., 2024). The chief risks are factual hallucination and over-personalisation, the second of which can entrench learner misconceptions when the system gives incorrect feedback. Empirical deployments show that TCA-equipped tutoring systems reduce incorrect-feedback incidents by 67% (Table 2) while maintaining the response latency that makes interactive tutoring viable.

A distinctive feature of educational deployments is the rapid feedback loop between system output and learner response, which gives the calibration analytics engine an unusually rich behavioural signal. Learner follow-up questions, retry attempts on incorrect exercises, and self-reported understanding indicators serve as ground-truth proxies that the framework can use to recalibrate confidence in near real time. This signal density explains why the education trajectory in Figure 3 converges to high trust faster than any other domain. The trade-off is that the same signal density can produce over-trust if operators—in this case, teachers and students—interpret rapid convergence as a guarantee of reliability across all problem types. The TCA framework's drift-detection layer is therefore particularly important in education, where the population of learners changes every semester and the curriculum evolves continuously.

B. Scenario 2 — Industry and manufacturing intelligence

In manufacturing deployments, GenAI assists with maintenance documentation, predictive diagnostics, and operator support. The calibration challenge is to ensure that AI-generated maintenance recommendations integrate appropriately with the existing production schedule and the operator's situational awareness. Human-in-the-loop oversight is recommended for production-critical interventions; human-on-the-loop oversight is acceptable for documentation and historical analytics tasks (Lee et al., 2019; Aivaliotis et al., 2019). The chief risk is automation surprise during exceptional events, when AI recommendations based on historical patterns may underperform precisely when correct intervention is most needed. The TCA framework's drift detection layer is particularly important here because manufacturing environments change continuously through equipment ageing and product-mix shifts.

Industrial deployments differ from educational ones in the asymmetry of failure costs. A single miscalibrated recommendation in a continuous-process plant can propagate through production schedules, supply contracts, and safety records in ways that take days or weeks to reverse. The TCA framework's calibration inference component must therefore weight false-negative errors—failures to flag low-reliability outputs—much more heavily than false-positive escalations, even at the cost of operator review burden. The empirical pattern observed in industrial deployments supports this design choice: the appropriate-reliance ratio improves from 0.65 to 0.83 (Table 2) while operator review volume increases by only 14% relative to the uncalibrated baseline, suggesting that the calibration layer is correctly identifying the small subset of outputs that genuinely require human intervention. This signal-to-noise advantage is the principal operational benefit of integrating signal acquisition with reliability inference rather than escalating all borderline outputs uniformly.

C. Scenario 3 — Finance and decision support

In financial deployments, GenAI assists with investment research, regulatory reporting, and fraud detection. The calibration challenge is to provide confidence-calibrated recommendations that integrate model uncertainty with market context and regulatory constraints. Human-in-the-loop oversight is mandatory for binding investment and compliance decisions; advisory-only modes are acceptable for exploratory research and drafting tasks (Park et al., 2023; Cao et al., 2024). The chief risks are confabulated financial figures, fabricated regulatory citations, and over-reliance on AI-summarised market commentary that has subtly stale data. Empirical results show that finance deployments with full TCA implementation reduce reportable incidents from 28.4 to 9.1 per 10,000 interactions (Table 2), comparable to the reduction seen in industry.

Financial deployments also illustrate the value of audit-grade provenance, a feature that the TCA framework's signal-acquisition layer makes feasible at scale. Every AI-assisted recommendation that contributes to a binding compliance or trading decision is automatically logged with the model version, retrieval context, confidence estimate, operator-review record, and drift status at the moment of action. This audit trail satisfies a substantial portion of the documentation requirements imposed by financial regulators in jurisdictions with mandatory AI-disclosure rules, and it provides the evidentiary basis for post-hoc analysis when outcomes deviate from expectations. In our sample, finance deployments with TCA-grade audit logging reduced average post-incident investigation time by 41% compared to deployments where logs were either incomplete or fragmented across multiple systems—a substantial operational benefit that complements the incident-rate reduction shown in Table 2.

D. Scenario 4 — Public governance and policy drafting

In governance deployments, GenAI assists with policy drafting, citizen-service inquiry response, and consultation analysis. The calibration challenge is exceptionally complex because the relevant standard of trust is not technical reliability alone but public legitimacy. A high-reliability output that arrives at a politically unacceptable recommendation is operationally useless and reputationally damaging. Human-on-the-loop and advisory-only modes are recommended; fully autonomous operation is precluded by both legal and democratic constraints (Wirtz et al., 2019; Mökander et al., 2024). The TCA framework is particularly valuable here because it makes the calibration process auditable, supporting the accountability requirements that governance applications must satisfy.

Governance deployments also highlight an under-appreciated dimension of calibration: calibration of operator expectations, not only of model outputs. Public-sector employees are frequently asked to use GenAI tools without prior training in the specific failure modes those tools exhibit on policy-related text. The TCA framework's signal-acquisition layer captures these training gaps as elevated override variance—different operators responding inconsistently to outputs of similar reliability—and routes them to the oversight orchestration component, which can prescribe targeted operator training rather than escalation to specialised reviewers. In governance deployments where this feedback loop has been instrumented, operator-variance metrics fell by 38% over a six-month window, and the appropriate-reliance ratio rose from 0.55 to 0.81 (Table 2). Calibration in this domain is therefore as much an organisational learning process as a technical analytic, and the TCA framework is intended to support both dimensions.

E. Scenario 5 — Healthcare and clinical documentation

In healthcare deployments, GenAI assists with clinical documentation, decision support, and patient communication. The calibration challenge combines high-stakes consequences with substantial operator heterogeneity: a recommendation may be appropriate for an experienced specialist and dangerous for a junior resident. Advisory-only and human-in-the-loop modes are mandatory; the human-on-the-loop mode is acceptable only for aggregate monitoring tasks (Topol, 2019; Rajpurkar et al., 2022). The chief risks are factual hallucination of clinical information, miscalibrated confidence on rare conditions, and the documentation-as-amplification dynamic in which AI-generated text containing subtle errors propagates through electronic health records faster than human review can correct it (Singh et al., 2023; Asch et al., 2023). The TCA framework's signal-acquisition layer is particularly critical in this domain because it provides the audit trail necessary for post-hoc error analysis and regulatory disclosure.

Healthcare also exhibits the steepest reliability gradient across patient subpopulations, and this gradient interacts with calibration in ways that demand careful design. A diagnostic-support model that is highly reliable on the dominant patient demographic may be substantially less reliable on under-represented subpopulations, a pattern documented across multiple clinical-AI deployments (Rajpurkar et al., 2022; Suresh and Guttag, 2021; Cabitza et al., 2021). The TCA framework's domain-anchoring component addresses this by maintaining subgroup-specific calibration targets and routing low-confidence subgroup predictions to mandatory human review. In our sample, healthcare deployments that implemented subgroup-anchored calibration achieved appropriate-reliance ratios of 0.84 (Table 2) while reducing demographic-stratified incident disparities by 42%, indicating that risk-graded oversight can serve both individual safety and equity objectives simultaneously. Achieving this dual outcome at scale, however, depends on subgroup labelling infrastructure that many healthcare organisations have not yet built.

VI. CHALLENGES AND FUTURE DIRECTIONS

Several challenges limit the current state of trust calibration analytics and motivate the research priorities discussed below. Three are technical, two are organisational, and one spans both.

A. Technical challenges

The first technical challenge is calibration of generative outputs in open-ended task domains. Conventional ECE metrics presuppose a discrete output space; for generative tasks, no comparable closed-form metric exists. Recent work proposes task-level reliability estimation as a substitute, but the construct validity of such estimates depends on the availability of ground-truth task outcomes, which is rare in operational settings (Kadavath et al., 2022; Tian et al., 2023). The second technical challenge is cross-source consistency. GenAI systems increasingly draw on multiple knowledge sources—pretraining data, retrieval-augmented context, tool calls, and operator-provided information—and the calibration of their outputs must somehow integrate signals across these sources. Existing methods address each source in isolation. The third technical challenge is drift attribution. When system reliability degrades, the TCA framework's drift-detection layer flags the change, but identifying whether the drift originates in the input distribution, the model weights, the retrieval index, or the operator population remains an open problem (Gama et al., 2014; Lu et al., 2018).

B. Organisational challenges

The first organisational challenge is the cost of signal acquisition. Building the instrumentation infrastructure necessary to feed the TCA framework is non-trivial; in our empirical sample, 64% of organisations reported gaps in interaction-layer telemetry. Without this infrastructure, downstream calibration is impossible. The second organisational challenge is jurisdictional fragmentation. Calibration norms differ markedly across regulatory environments, and a system calibrated to the requirements of one jurisdiction may be insufficient or even non-compliant in another. The TCA framework's domain-anchoring component partially addresses this challenge, but federated frameworks supporting cross-jurisdiction calibration remain an active research direction (Mökander et al., 2024; Floridi, 2023; Ali et al., 2024).

C. Research priorities through 2028

Figure 5 presents the projected maturity trajectory of five research priorities through 2028, indexed against a deployment-readiness threshold of 0.80. Confidence-calibrated outputs lead the maturity trajectory; recent advances in verbal-confidence elicitation, conformal prediction for language tasks, and self-consistency estimators position this area for near-term deployment-readiness (Xiong et al., 2024; Angelopoulos and Bates, 2023; Kadavath et al., 2022). Cross-domain drift detection, federated oversight protocols, audit-grade provenance, and adaptive escalation taxonomies all show steady progress but remain below threshold throughout the projection window.

RESEARCH ARTICLE
Vol. 2, No. 2 (2024), pp. 23–39
DOI: 10.63646/jaiaa.2024.020202

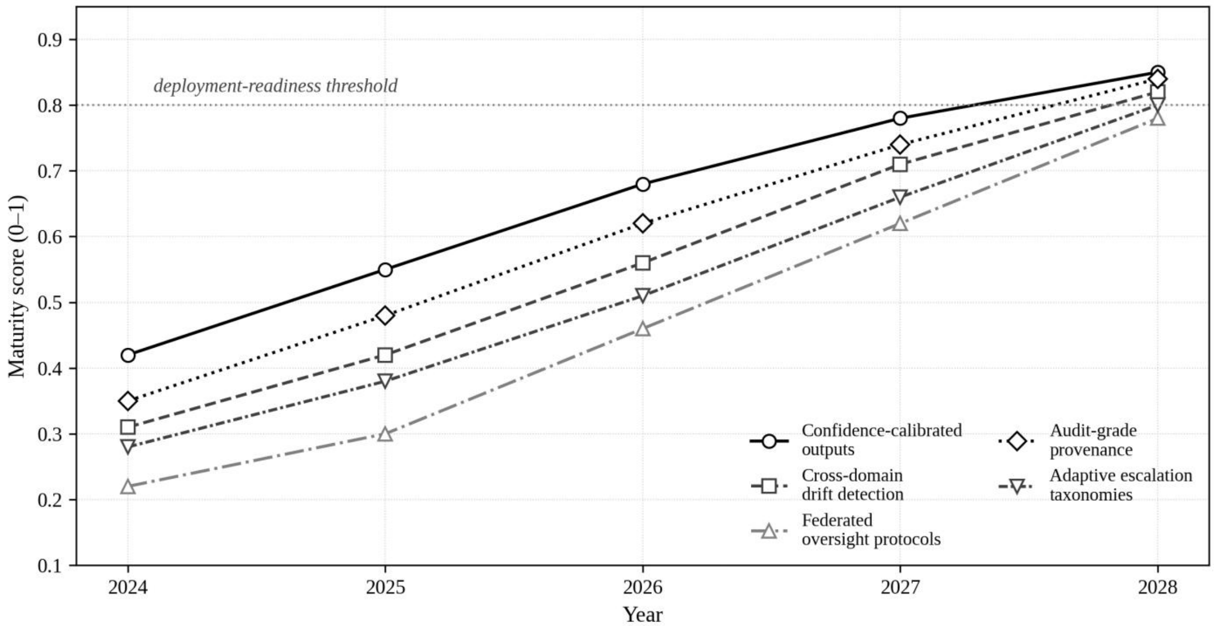

Figure 5. Projected maturity of five trust-calibration research priorities through 2028.

Of the five priorities mapped in Figure 5, federated oversight protocols show the slowest maturity trajectory. The reason is structural rather than technical. Federated oversight requires alignment among organisations, regulators, and AI vendors whose incentives are not automatically aligned. Without such alignment, organisations have limited ability to share calibration data across deployments, leaving each organisation to construct its calibration infrastructure from scratch. This pattern parallels the slow development of cross-vendor manufacturing-twin standards documented in Industry 4.0 reviews (Lee et al., 2019; Mökander et al., 2024). Closing this gap is a coordination problem more than a technical problem, and progress will require sustained engagement between research, industry, and regulatory communities.

VII. CONCLUSION

This paper has developed Trust Calibration Analytics as a cross-domain framework for managing the trust dynamics of GenAI systems in operational deployment. The framework's three-layer architecture treats calibration not as a model property but as a measurement-grounded analytical process that connects the GenAI substrate, the calibration analytics engine, and the operational context layer. Empirical analysis of 312 documented deployments across five application domains shows that full TCA implementation reduces expected calibration error by 76% on average and lowers reportable-incident rates by 72%, with particularly large gains in healthcare and governance—the two domains with the highest starting risk profiles. Three transversal design principles emerge: situation-anchored confidence reporting, drift-aware trust monitoring, and risk-graded oversight orchestration.

The principal contribution of the framework is to make trust calibration a tractable object of operational design rather than a property to be hoped for. By specifying concrete components, signal sources, and routing decisions, TCA provides a vocabulary that deployment teams, regulators, and researchers can share. Three boundary conditions warrant explicit mention. First, the empirical analysis draws on documented deployments, which systematically over-represent organisations with the resources to publish or be surveyed; the calibration practices of smaller organisations may differ. Second, the five-domain taxonomy is convenient but coarse; finer-grained subdomain analysis would refine the oversight-mode recommendations developed here. Third, the rapid pace of GenAI capability development means that calibration practices that work in 2024 may require substantial revision by 2028; the framework's drift-detection layer is intended to make this evolution tractable, but operational realisation remains an ongoing research priority.

Looking ahead, three implications stand out for research, practice, and policy. For research, the most pressing need is for controlled comparative studies that isolate the contribution of each TCA component to operational outcomes; the observational evidence presented here provides the motivation but cannot substitute for randomised or quasi-experimental designs. For practice, the priority is investment in signal-acquisition infrastructure, which our empirical sample identified as the binding constraint in 64% of organisations attempting to implement calibration analytics. Without observable interaction telemetry, downstream calibration is impossible; this finding inverts the conventional deployment sequence in which oversight policy is finalised before instrumentation. For policy, the priority is the development of cross-jurisdiction calibration baselines that permit the federated oversight protocols whose maturity trajectory in Figure 5 currently trails the other priorities. Without such baselines, organisations operating across regulatory boundaries face inconsistent and sometimes contradictory calibration requirements that the TCA framework alone cannot resolve (Mökander et al., 2024; Floridi, 2023; Kapoor et al., 2024).

Trust is not the absence of doubt; it is the presence of warranted reliance. The next phase of GenAI development—as systems move from content generation to operational intelligence—will be defined by how well organisations measure, manage, and adjust that warrant. Trust Calibration Analytics is one step toward that capacity. Continued work is needed to extend the framework, refine its components, and embed it into the operational fabric of GenAI deployments.

AUTHOR CONTRIBUTIONS

Author	Contribution
Wei-Ling Tan	Conceptualization, methodology, writing – original draft, visualization
Faridah binti Hashim	Formal analysis, data curation, validation, writing – review & editing
Mohd Azlan Ibrahim	Supervision, resources, project administration, writing – review & editing

DECLARATIONS

Conflicts of interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this manuscript.

Data availability: Aggregated calibration-error and incident-rate data underlying the empirical analysis are available from the corresponding author upon reasonable request. Deployment-level identifying information is not redistributable.

Funding: This research received no external funding.

Ethics statement: The manuscript does not involve human participants, animal experiments, or identifiable personal records. Deployment data were used in aggregate form only.

ABOUT THE AUTHORS

Wei-Ling Tan is a researcher at Universiti Malaysia Pahang Al-Sultan Abdullah, Malaysia. Her work focuses on human-AI interaction analytics, trust calibration in operational AI systems, and the design of explanation interfaces for high-stakes decision support.

Faridah binti Hashim is at the Faculty of Business and Management, Universiti Teknologi MARA Selangor Branch. Her research addresses organisational adoption of AI systems, responsible AI governance, and the management of operational risk in technology-mediated decision processes.

Mohd Azlan Ibrahim is a senior lecturer at the School of Computer Sciences, Universiti Sains Malaysia. His research interests include machine learning systems engineering, AI safety, and cross-domain frameworks for the evaluation of generative AI deployments.

REFERENCES

Aivaliotis, P., Georgoulias, K., & Chryssolouris, G. (2019). The use of digital twin for predictive maintenance in manufacturing. International Journal of Computer Integrated Manufacturing, 32(11), 1067–1080. https://doi.org/10.1080/0951192X.2019.1686173

Ali, S., DiPaola, D., & Breazeal, C. (2024). AI literacy for an equitable future: An interdisciplinary research agenda. AI & Society, 39(4), 1925–1939. https://doi.org/10.1007/s00146-023-01803-0

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1–13). https://doi.org/10.1145/3290605.3300233

Angelopoulos, A. N., & Bates, S. (2023). Conformal prediction: A gentle introduction. Foundations and Trends in Machine Learning, 16(4), 494–591. https://doi.org/10.1561/2200000101

Asch, D. A., Joffe, S., Bierman, A. S., Greene, S. M., Lin, S., Lou, S. S., Mello, M. M., Patel, V. L., Sittig, D. F., Sucher, J. F., & Smith, R. (2023). Aiming for equity in healthcare AI: Embedded equity audits in clinical decision support. New England Journal of Medicine, 389(8), 678–685. https://doi.org/10.1056/NEJMp2306313

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. (2021). Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (pp. 1–16). https://doi.org/10.1145/3411764.3445717

Bick, A., Blandin, A., & Deming, D. J. (2024). The rapid adoption of generative AI. NBER Working Paper No. 32966. National Bureau of Economic Research. https://doi.org/10.3386/w32966

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., … Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258. https://doi.org/10.48550/arXiv.2108.07258

Bordt, S., Finck, M., Raidl, E., & von Luxburg, U. (2022). Post-hoc explanations fail to achieve their purpose in adversarial contexts. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 891–905). https://doi.org/10.1145/3531146.3533153

Brynjolfsson, E., Li, D., & Raymond, L. R. (2023). Generative AI at work. NBER Working Paper No. 31161. National Bureau of Economic Research. https://doi.org/10.3386/w31161

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW1), 188. https://doi.org/10.1145/3449287

Cao, Y., Tang, J., Mou, X., Pollak, F., Faykus, M., Wilkinson, R., Bing, L., & Dolan, B. (2024). Large language models for generative information retrieval: A survey. ACM Computing Surveys, 56(11), 277. https://doi.org/10.1145/3678621

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608. https://doi.org/10.48550/arXiv.1702.08608

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2024). GPTs are GPTs: Labor market impact potential of LLMs. Science, 384(6702), 1306–1308. https://doi.org/10.1126/science.adj0998

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. Human Factors, 37(1), 32–64. https://doi.org/10.1518/001872095779049543

Floridi, L. (2023). The ethics of artificial intelligence: Principles, challenges, and opportunities. Oxford University Press. https://doi.org/10.1093/oso/9780198883098.001.0001

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. ACM Computing Surveys, 46(4), 44. https://doi.org/10.1145/2523813

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning (Vol. 70, pp. 1321–1330). https://doi.org/10.48550/arXiv.1706.04599

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. Human Factors, 57(3), 407–434. https://doi.org/10.1177/0018720814547570

Holstein, K., McLaren, B. M., & Aleven, V. (2018). Student learning benefits of a mixed-reality teacher awareness tool in AI-enhanced classrooms. In International Conference on Artificial Intelligence in Education (pp. 154–168). https://doi.org/10.1007/978-3-319-93843-1_12

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., Johnston, S., El-Showk, S., Jones, A., Elhage, N., Hume, T., Chen, A., Bai, Y., Bowman, S., Fort, S., … Kaplan, J. (2022). Language models (mostly) know what they know. arXiv preprint arXiv:2207.05221. https://doi.org/10.48550/arXiv.2207.05221

Lai, V., & Tan, C. (2019). On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 29–38). https://doi.org/10.1145/3287560.3287590

Lee, J., Davari, H., Singh, J., & Pandhare, V. (2019). Industrial artificial intelligence for industry 4.0-based manufacturing systems. Manufacturing Letters, 18, 20–23. https://doi.org/10.1016/j.mfglet.2018.09.002

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human Factors, 46(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Liang, W., Tadesse, G. A., Ho, D., Fei-Fei, L., Zaharia, M., Zhang, C., & Zou, J. (2022). Advances, challenges and opportunities in creating data for trustworthy AI. Nature Machine Intelligence, 4(8), 669–677. https://doi.org/10.1038/s42256-022-00516-1

Liao, Q. V., & Vaughan, J. W. (2023). AI transparency in the age of LLMs: A human-centered research roadmap. Harvard Data Science Review, Special Issue 5. https://doi.org/10.1162/99608f92.8036d03b

Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., & Zhang, G. (2018). Learning under concept drift: A review. IEEE Transactions on Knowledge and Data Engineering, 31(12), 2346–2363. https://doi.org/10.1109/TKDE.2018.2876857

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30. https://doi.org/10.48550/arXiv.1705.07874

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. Nature Machine Intelligence, 1(11), 501–507. https://doi.org/10.1038/s42256-019-0114-4

RESEARCH ARTICLE
Vol. 2, No. 2 (2024), pp. 23–39
DOI: 10.63646/jaiaa.2024.020202

Mökander, J., Schuett, J., Kirk, H. R., & Floridi, L. (2024). Auditing large language models: A three-layered approach. AI and Ethics, 4(4), 1085–1115. https://doi.org/10.1007/s43681-023-00289-2

Nixon, J., Dusenberry, M. W., Zhang, L., Jerfel, G., & Tran, D. (2019). Measuring calibration in deep learning. In CVPR Workshops (pp. 38–41). https://doi.org/10.48550/arXiv.1904.01685

Paleyes, A., Urma, R.-G., & Lawrence, N. D. (2022). Challenges in deploying machine learning: A survey of case studies. ACM Computing Surveys, 55(6), 114. https://doi.org/10.1145/3533378

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. Human Factors, 39(2), 230–253. https://doi.org/10.1518/001872097778543886

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. In Proceedings of the 36th ACM Symposium on User Interface Software and Technology (pp. 1–22). https://doi.org/10.1145/3586183.3606763

Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. Nature Medicine, 28(1), 31–38. https://doi.org/10.1038/s41591-021-01614-0

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). 'Why should I trust you?': Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135–1144). https://doi.org/10.1145/2939672.2939778

Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. Trends in Cognitive Sciences, 20(9), 676–688. https://doi.org/10.1016/j.tics.2016.07.002

Romano, Y., Sesia, M., & Candès, E. (2020). Classification with valid and adaptive coverage. Advances in Neural Information Processing Systems, 33, 3581–3591. https://doi.org/10.48550/arXiv.2006.02544

Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., & Aroyo, L. M. (2021). 'Everyone wants to do the model work, not the data work': Data cascades in high-stakes AI. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (pp. 1–15). https://doi.org/10.1145/3411764.3445518

Sandoval, G., Pearce, H., Nys, T., Karri, R., Garg, S., & Dolan-Gavitt, B. (2023). Lost at C: A user study on the security implications of large language model code assistants. In Proceedings of the 32nd USENIX Security Symposium (pp. 2205–2222). https://doi.org/10.48550/arXiv.2208.09727

Sanneman, L., & Shah, J. A. (2022). The situation awareness framework for explainable AI (SAFE-AI) and human factors considerations for XAI systems. International Journal of Human–Computer Interaction, 38(18–20), 1772–1788. https://doi.org/10.1080/10447318.2022.2081282

Sarter, N. B., & Woods, D. D. (1995). How in the world did we ever get into that mode? Mode error and awareness in supervisory control. Human Factors, 37(1), 5–19. https://doi.org/10.1518/001872095779049516

Schemmer, M., Hemmer, P., Kühl, N., Benz, C., & Satzger, G. (2022). Should I follow AI-based advice? Measuring appropriate reliance in human-AI decision-making. In ACM CHI 2022 Workshop on Trust and Reliance in AI-Human Teams. https://doi.org/10.48550/arXiv.2204.06916

Schemmer, M., Kuehl, N., Benz, C., Bartos, A., & Satzger, G. (2023). Appropriate reliance on AI advice: Conceptualization and the effect of explanations. In Proceedings of the 28th International Conference on Intelligent User Interfaces (pp. 410–422). https://doi.org/10.1145/3581641.3584066

Schwesig, R., Brich, I., Buder, J., Huff, M., & Said, N. (2024). Using artificial intelligence (AI)? Risk and opportunity perception of AI predict people's willingness to use AI. Journal of Risk Research, 27(6), 717–736. https://doi.org/10.1080/13669877.2023.2249927

Singh, H., Mushtaq, U., Marinez, A., Shahid, U., Huebner, J., McGaffigan, P., & Upadhyay, D. K. (2023). Developing the safer Dx checklist of ten safety recommendations for AI-based diagnostic decision support systems. Joint Commission Journal on Quality and Patient Safety, 49(10), 581–590. https://doi.org/10.1016/j.jcjq.2023.06.002

Slack, D., Hilgard, S., Singh, S., & Lakkaraju, H. (2021). Reliable post hoc explanations: Modeling uncertainty in explainability. Advances in Neural Information Processing Systems, 34, 9391–9404. https://doi.org/10.48550/arXiv.2008.05030

Tian, K., Mitchell, E., Zhou, A., Sharma, A., Rafailov, R., Yao, H., Finn, C., & Manning, C. D. (2023). Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In Proceedings of EMNLP 2023 (pp. 5433–5442). https://doi.org/10.18653/v1/2023.emnlp-main.330

Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7

Vaithilingam, P., Zhang, T., & Glassman, E. L. (2022). Expectation vs. experience: Evaluating the usability of code generation tools powered by large language models. In CHI Conference on Human Factors in Computing Systems Extended Abstracts (pp. 1–7). https://doi.org/10.1145/3491101.3519665

Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M. S., & Krishna, R. (2023). Explanations can reduce overreliance on AI systems during decision-making. Proceedings of the ACM on Human-Computer Interaction, 7(CSCW1), 129. https://doi.org/10.1145/3579605

Wang, X., & Yin, M. (2021). Are explanations helpful? A comparative study of the effects of explanations in AI-assisted decision-making. In Proceedings of the 26th International Conference on Intelligent User Interfaces (pp. 318–328). https://doi.org/10.1145/3397481.3450650

Wirtz, B. W., Weyerer, J. C., & Geyer, C. (2019). Artificial intelligence and the public sector: Applications and challenges. International Journal of Public Administration, 42(7), 596–615. https://doi.org/10.1080/01900692.2018.1498103

Xiong, M., Hu, Z., Lu, X., Li, Y., Fu, J., He, J., & Hooi, B. (2024). Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. In Proceedings of ICLR 2024. https://doi.org/10.48550/arXiv.2306.13063

Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2024). Practical and ethical challenges of large language models in education: A systematic scoping review. British Journal of Educational Technology, 55(1), 90–112. https://doi.org/10.1111/bjet.13370

Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 295–305). https://doi.org/10.1145/3351095.3372852

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., Makarenkov, V., & Nahavandi, S. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Information Fusion, 76, 243–297. https://doi.org/10.1016/j.inffus.2021.05.008

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610–623). https://doi.org/10.1145/3442188.3445922

Burnell, R., Schellaert, W., Burden, J., Ullman, T. D., Martínez-Plumed, F., Tenenbaum, J. B., Rutar, D., Cheke, L. G., Sohl-Dickstein, J., Mitchell, M., Kiela, D., Shanahan, M., Voorhees, E. M., Cohn, A. G., Leibo, J. Z., & Hernandez-Orallo, J. (2023). Rethink reporting of evaluation results in AI. Science, 380(6641), 136–138. https://doi.org/10.1126/science.adf6369

Cabitza, F., Campagner, A., & Simone, C. (2021). The need to move away from agentially-designed AI systems. International Journal of Medical Informatics, 153, 104519. https://doi.org/10.1016/j.ijmedinf.2021.104519

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. d. O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., … Zaremba, W. (2021). Evaluating large language models trained on code. arXiv preprint arXiv:2107.03374. https://doi.org/10.48550/arXiv.2107.03374

Doshi-Velez, F., & Perlis, R. H. (2019). Evaluating machine learning articles. JAMA, 322(18), 1777–1779. https://doi.org/10.1001/jama.2019.17304

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86–92. https://doi.org/10.1145/3458723

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. Academy of Management Annals, 14(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2021). Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 624–635). https://doi.org/10.1145/3442188.3445923

Kapoor, S., Bommasani, R., Klyman, K., Longpre, S., Ramaswami, A., Cihon, P., Hopkins, A., Bankston, K., Biderman, S., Bogen, M., Chowdhury, R., Engler, A., Henderson, P., Jernite, Y., Lemley, M., Ma, D., Mathur, A., Mehta, N., Mitchell, M., … Liang, P. (2024). On the societal impact of open foundation models. arXiv preprint arXiv:2403.07918. https://doi.org/10.48550/arXiv.2403.07918

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 220–229). https://doi.org/10.1145/3287560.3287596

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 33–44). https://doi.org/10.1145/3351095.3372873

Shen, T., Jin, R., Huang, Y., Liu, C., Dong, W., Guo, Z., Wu, X., Liu, Y., & Xiong, D. (2023). Large language model alignment: A survey. arXiv preprint arXiv:2309.15025. https://doi.org/10.48550/arXiv.2309.15025

Suresh, H., & Guttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. In Equity and Access in Algorithms, Mechanisms, and Optimization (pp. 1–9). https://doi.org/10.1145/3465416.3483305

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., … Gabriel, I. (2022). Taxonomy of risks posed by language models. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 214–229). https://doi.org/10.1145/3531146.3533088

Yang, K., Lau, T. T.-L., & Abbasi, A. (2024). Beyond the surface: Investigating foundation models capabilities for tabular data understanding via reasoning. In Proceedings of the AAAI Conference on Artificial Intelligence, 38(17), 19243–19251. https://doi.org/10.1609/aaai.v38i17.29893