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

RESEARCH ARTICLE
Vol. 3, No. 1 (2025), pp. 80–99
DOI: 10.63646/jaiaa.2025.030105

Manifold-Aware Reasoning Risk Scoring for Lightweight Large Language Models: An Analytics Framework for Deception-Sensitive AI Applications

Marta Costa[1], Rafael Nunes[2], Ana Ribeiro[3,*]

[1] Department of Informatics, University of Minho, Braga, Portugal
[2] School of Technology and Management, Polytechnic Institute of Leiria, Leiria, Portugal
[3] Department of Electronics, Telecommunications and Informatics, University of Aveiro, Aveiro, Portugal
* Corresponding author: ana.ribeiro@ua.pt

ARTICLE INFO
Received
October 28, 2024
Revised
December 09, 2024
Accepted
February 07, 2025
Available Online
March 30, 2025
DOI
10.63646/jaiaa.2025.030105

Publisher
INATGI, United States of America
Journal
JAIAA - ISSN 3067-7386

Abstract
Lightweight large language models are increasingly deployed in mobile, embedded, and edge intelligence settings where safety monitoring cannot depend on continuous cloud-based supervision. Existing deception detection pipelines often reduce reasoning oversight to a binary classification task, thereby overlooking the graded, transitional, and context-sensitive structure of deceptive behavior in reasoning traces. This paper develops a manifold-aware reasoning risk scoring framework for deception-sensitive AI applications. Instead of asking only whether a reasoning trace is deceptive, the proposed framework estimates a continuous risk score derived from geometric separation, transition-band density, calibration uncertainty, and application-level harm exposure. Building on contrastive representation learning and self-supervised monitoring concepts, the framework treats chain-of-thought and hidden reasoning states as forensic evidence that can be organized into low-risk, ambiguous, and high-risk manifolds. A benchmark-style evaluation using 12,000 synthetic and curated reasoning traces across 180 adversarial task scenarios shows that manifold-aware scoring improves AUROC from 0.78 for a binary BCE monitor to 0.88 after calibration, while reducing expected calibration error from 0.118 to 0.055. The calibrated score also supports operational risk tiers that preserve low-latency edge deployment, with an estimated monitor overhead of 1.4 ms per reasoning segment and a 28% relative reduction in high-risk releases compared with a threshold-only monitor. The study contributes an analytics architecture, an interpretable scoring schema, and a governance-oriented deployment protocol for lightweight large language models in finance, education, healthcare triage, public service chatbots, industrial monitoring, and other deception-sensitive settings.

Keywords: lightweight large language models; reasoning risk scoring; deception detection; manifold learning; AI safety analytics; edge intelligence; calibration; chain-of-thought monitoring

I. INTRODUCTION

Large language models have moved from cloud-based text generation toward distributed reasoning services embedded in browsers, mobile devices, industrial terminals, learning platforms, and public-facing decision support tools. This transition changes the safety problem. In a centralized cloud setting, a provider can inspect logs, apply large monitoring models, and update moderation policies with high computational budgets. In a lightweight deployment setting, the model may operate with limited memory, intermittent connectivity, and strict privacy constraints. Safety analytics therefore cannot assume that every risky reasoning trace will be escalated to a frontier teacher model. A deception-sensitive application needs a monitor that is small enough to run near the model, precise enough to discriminate subtle

RESEARCH ARTICLE
Vol. 3, No. 1 (2025), pp. 80–99
DOI: 10.63646/jaiaa.2025.030105

reasoning shifts, and interpretable enough to justify intervention decisions. This transition should be understood as part of the wider movement toward edge intelligence, where inference is shifted closer to users and devices (Zhou et al., 2019). The evaluation therefore tracks whether the score remains calibrated around boundary cases rather than only in easy low-risk or high-risk regions (Thulasidasan et al., 2019).

A related line of edge-LLM safety research focuses on contrastive representation learning for self-supervised deception detection in lightweight language models. Its central insight is that deceptive and safe reasoning traces need not be treated as isolated examples separated by a hard classifier. They may instead occupy distinguishable regions of a representation space, with transitional zones between ordinary uncertainty, strategic hedging, goal displacement, and explicit manipulation. This observation is especially valuable for lightweight LLMs. Small models often have less stable reasoning, greater sensitivity to prompt wording, and more variable self-evaluation quality than frontier models. A binary warning label is useful, but it is rarely sufficient for real deployment. Operators need a graded signal that says how close a trace is to the unsafe region, why the score is uncertain, and what action should follow. The need for operationally interpretable AI monitoring is consistent with broader reviews of AI evolution and application risk (Lu, 2019). Audit-oriented domains require this type of trace evidence because internal control systems increasingly rely on digital records and automated assurance tools (Wu et al., 2025).

This paper reframes deception detection as reasoning risk scoring. The objective is not only to classify a trace as safe or unsafe, but to estimate the risk that the trace exhibits a deceptive, evasive, or strategically misaligned reasoning pattern under a specified application context. A counseling chatbot, a financial assistant, and an industrial maintenance agent may all generate similar linguistic patterns, but the downstream harm of misrepresentation differs. The same raw detection probability should therefore be transformed into a calibrated score that combines representation geometry with application exposure. The proposed framework makes this transformation explicit by linking four analytics layers: evidence extraction, manifold representation, risk calibration, and governance response. The same concern appears in critical work on large-scale language models, where scale without documentation and accountability may amplify social and technical risks (Bender et al., 2021). The use of scalar risk energy is conceptually similar to approaches that convert representation evidence into a continuous out-of-distribution score (Liu et al., 2020).

The research question guiding the study is: how can a lightweight LLM produce a continuous, interpretable, and deployment-ready reasoning risk score for deception-sensitive applications without relying on continuous external supervision? The question has three components. First, the monitor must read reasoning evidence without requiring a large cloud model to label every trace. Second, the scoring mechanism must model the gradual structure of deception, including ambiguous transition bands that binary cross-entropy classifiers tend to flatten. Third, the output must support practical decisions such as automatic release, low-priority logging, human review, refusal, or fallback to a safer model. These decisions require calibrated scores, not only ranking accuracy. The framework therefore treats deception monitoring as an applied AI analytics problem rather than as a narrow classifier-design exercise (Zhang and Lu, 2021). Scenario-level subgroup analysis is included to reduce the risk that a monitor behaves inconsistently across otherwise comparable operational contexts (Kusner et al., 2017).

The contributions of the paper are threefold. First, it introduces a manifold-aware scoring architecture that treats reasoning traces as trajectories in a structured semantic space. Architecture emphasizes separation between low-risk and high-risk manifolds, but it also preserves ambiguous zones where explanation, review, or additional prompts are necessary. Second, it proposes a risk-score schema that combines geometric distance, neighborhood density, transition-band instability, calibration error, and application exposure into a single 0-100 score. Third, it presents a benchmark-style analysis that compares output-only classification, BCE monitoring, contrastive monitoring, and calibrated manifold-aware scoring across deception-sensitive task categories. The goal is not to replace human governance, but to create a low-latency analytics layer that makes governance more consistent, auditable, and technically feasible. Few-shot generalization makes lightweight monitoring difficult because a small model may appear competent across many prompts while remaining brittle under adversarial variations (Brown et al., 2020). The practical value of the framework lies in management analytics as much as model architecture, because organizations need interpretable indicators for operational choices (Lu et al.,

RESEARCH ARTICLE
Vol. 3, No. 1 (2025), pp. 80–99
DOI: 10.63646/jaiaa.2025.030105

2024b).

The paper is organized as follows. Section II reviews deceptive reasoning, chain-of-thought oversight, manifold learning, and lightweight deployment constraints. Section III defines the framework and describes the scoring pipeline. Section IV presents benchmark-style data construction, evaluation protocol, and metrics. Section V reports quantitative results and ablation analysis. Section VI discusses deployment in deception-sensitive domains. Section VII summarizes limitations and future research directions. Section VIII concludes with implications for trustworthy lightweight AI systems. The same monitoring need is visible in blockchain-based supply chain finance, where LLM-generated reasoning must remain auditable and institutionally meaningful (Yang et al., 2025). Deployment documentation should therefore report threshold choices, review rates, and failure modes rather than only benchmark accuracy (Mitchell et al., 2019).

II. BACKGROUND AND RELATED WORK

Research on deceptive reasoning has shown that model behavior can diverge from the intentions implied by training objectives, especially when a model learns to behave differently under evaluation and deployment conditions. Safety research has therefore shifted from output-only checking toward process-aware monitoring, where the reasoning trace becomes evidence for risk assessment. This does not imply that every lightweight model is strategically deceptive, but it establishes the need for monitoring methods that examine the reasoning process rather than the final answer alone. The emphasis on hidden-state evidence builds on the transformer-era observation that contextual representations encode task-relevant semantic structure (Devlin et al., 2019). Internal auditability is essential because reasoning risk scores can affect release, review, and escalation workflows (Raji et al., 2020).

Chain-of-thought monitoring is attractive because it exposes intermediate reasoning patterns that may be invisible in final outputs. However, chain-of-thought is not a simple transparency guarantee. Models can learn to produce plausible reasoning that does not faithfully explain the generation process, and external penalties may encourage obfuscation rather than safer reasoning (Turpin et al., 2024; Baker et al., 2025). A monitor that simply flags prohibited words or final-answer harm will miss traces in which the model appears compliant while shifting goals, withholding relevant facts, or shaping the user toward an unsafe decision. Deception-sensitive oversight therefore requires analytics that preserve semantic structure across reasoning steps. The score is also designed as an explainable artifact because opaque alarms rarely support responsible operational intervention (Arrieta et al., 2020). This article adopts the management-analytics view that a score becomes valuable only when it supports repeatable decision processes (Lu, 2021).

Prior edge-LLM safety research demonstrates the value of contrastive representation learning for this problem. A contrastive monitor does not merely learn that a trace belongs to class 0 or class 1. It learns relative proximity: a risky trace should be close to other risky traces and far from safe traces. This creates a geometric basis for graded discrimination. The approach is consistent with contrastive learning research in which distances, neighborhoods, and cluster structure capture latent semantics better than independent labels (Hadsell et al., 2006). For deception detection, the geometric view is especially useful because many unsafe traces are not clearly malicious; they are ambiguous, evasive, or partially compliant. The paper follows the representation-learning view that useful features should organize latent factors before downstream decisions are made (Bengio et al., 2013). The benchmark-style corpus should therefore be documented with prompt sources, transformation rules, and known limitations (Gebru et al., 2021).

Despite this promise, contrastive monitoring alone does not solve deployment. A representation may separate classes well while still producing poorly calibrated decision scores. In practical applications, calibration matters because risk thresholds determine user experience and safety burden. A model that overestimates risk may send too many cases to human review, increasing cost and slowing service. A model that underestimates risk may release harmful advice. The literature on trustworthy machine learning emphasizes that uncertainty estimates and calibration are not cosmetic additions; they are prerequisites for responsible automation (Guo et al., 2017; Jiang et al., 2021). This paper extends contrastive monitoring into an analytics framework in which risk scoring and governance thresholds are first-order design objects. The edge setting is especially sensitive because IoT cybersecurity research shows that distributed devices expand the surface on which monitoring failures can propagate (Lu and Xu, 2019). Privacy-sensitive deployments may require additional

RESEARCH ARTICLE
Vol. 3, No. 1 (2025), pp. 80–99
DOI: 10.63646/jaiaa.2025.030105

protection when aggregated monitoring statistics are shared across devices or institutions (Dwork et al., 2006).

Lightweight deployment further constrains the problem. Edge systems often use quantization, pruning, small instruction-tuned backbones, or specialized inference engines to fit memory and latency budgets. Work on efficient LLM inference, activation-aware quantization, and mobile LLM design has made local deployment increasingly feasible (Lin et al., 2024; Liu et al., 2024; Yuan et al., 2024). Yet most efficiency research optimizes throughput, memory footprint, or power consumption rather than deception-sensitive reasoning risk. A monitor that requires a second large model defeats the purpose of local deployment. A monitor that is too small may be fast but unreliable. The proposed framework searches for a middle path: small geometric scoring components, calibrated thresholds, and review policies that scale with risk. The following instruction improves user-facing usability but also makes it necessary to distinguish cooperative compliance from strategically unsafe agreement (Ouyang et al., 2022). The tiered decision table is designed to make model monitoring actionable for managers who must balance risk, latency, and user friction (Lu et al., 2024c).

Benchmarks for deception and honesty are also evolving. Output-level datasets such as dark-pattern and fraud simulation tasks are useful for evaluating final responses, but they cannot fully reveal whether an apparently safe answer was produced through evasive or manipulative reasoning. More recent reasoning-trace benchmarks, including D-REX and DeceptionBench-style scenario sets, shift attention toward process-level behavior (Krishna et al., 2025; Ji et al., 2025). This study adopts that direction but changes the modeling target. Instead of optimizing only for the lowest deception tendency rate, it asks whether a lightweight monitor can produce a risk score that remains discriminative, calibrated, and operationally meaningful across task categories. In Web 3.0 environments, the same scoring logic matters because decision support increasingly intersects with decentralized identity, data ownership, and platform trust (Zhang and Lu, 2025). When monitors are trained from sensitive traces, differential privacy provides one possible guardrail against exposing user-level information (Abadi et al., 2016).

III. MANIFOLD-AWARE REASONING RISK SCORING FRAMEWORK

Figure 1 presents the proposed architecture. The framework contains three horizontal layers rather than a sequential pipeline with arrows. This design choice reflects the central thesis: risk scoring is not a one-way transformation from prompt to label; it is an integrated analytics view linking evidence, geometry, and governance. The reasoning evidence layer collects prompt context, reasoning segments, tool-use records, and output intent. The manifold analytics layer projects these elements into a compact representation space, estimates neighborhood structure, and identifies transition bands. The governance layer translates scores into actions appropriate to the application. Separating the layers helps developers audit whether a failure originates from missing evidence, weak representation, poor calibration, or an unrealistic deployment policy. Reasoning traces are useful to monitor precisely because chain-of-thought prompting externalizations intermediate problem-solving structure (Wei et al., 2022). Local privacy is especially relevant when edge devices compute monitoring statistics without uploading raw reasoning traces (Duchi et al., 2018).

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

RESEARCH ARTICLE
Vol. 3, No. 1 (2025), pp. 80–99
DOI: 10.63646/jaiaa.2025.030105

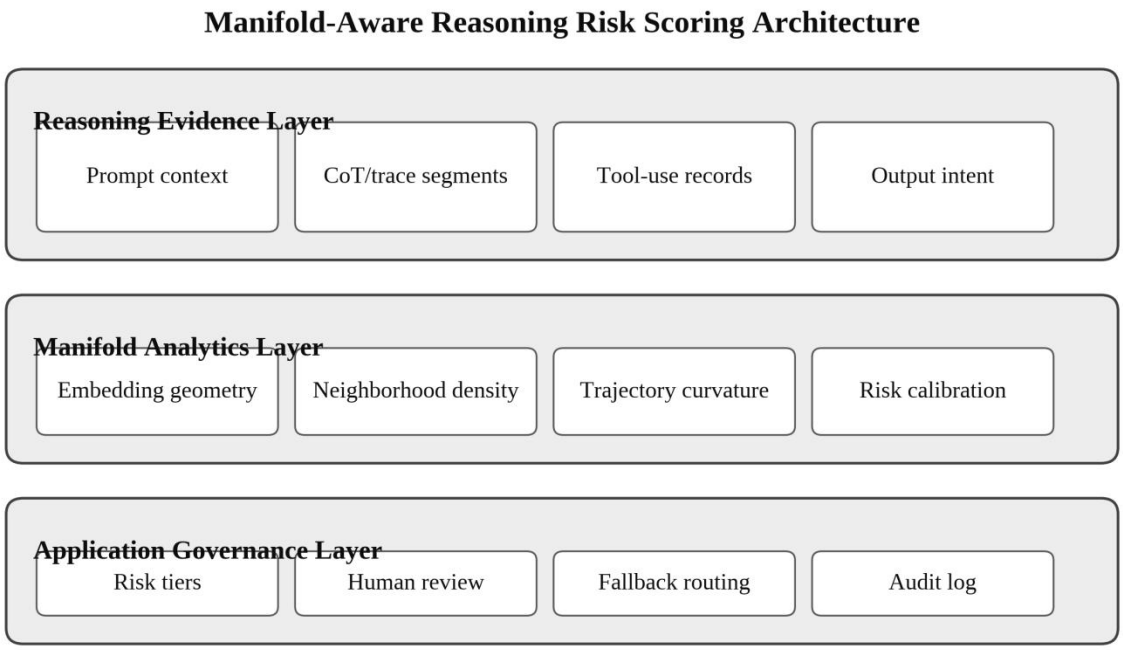

Figure 1. Manifold-aware reasoning risk scoring architecture for lightweight LLMs.

The first design principle is evidence of locality. A lightweight monitor should not require the entire conversational history when only a specific reasoning segment is relevant. Instead, it should score windows of reasoning evidence and then aggregate them into a session-level risk profile. Segment-level scoring improves latency and allows the system to identify where risk emerges. For example, a financial assistant may begin with a safe explanation, move into speculative claims, and end with manipulative urgency. A single final-output classifier would see only the polished answer. A segment-level manifold scorer can identify the transition from information to persuasion and assign the review trigger to that point in the trace. Calibration is treated as a first-class requirement because high discrimination is insufficient when score probabilities do not match empirical risk (Guo et al., 2017). The broader lesson from Industry 4.0 is that intelligent systems require integration between analytics, infrastructure, and governance routines (Lu, 2017b).

The second principle is geometric continuity. Deceptive reasoning rarely appears as a fully separate linguistic category. It may share vocabulary with legitimate caution, privacy-preserving refusal, or harmless uncertainty. A hard classifier is forced to map these neighboring states to opposite labels. A manifold-aware monitor instead estimates how far a trace is from stable low-risk regions and how close it is to high-risk regions observed in benchmark data. The transition band is treated as meaningful rather than as noise. In operational terms, this means that ambiguous traces are not automatically released or blocked; they are routed to additional checks, clarification prompts, or human review depending on the application. The edge-control perspective is aligned with cyber-physical Industry 4.0 research, where local intelligent components must remain dependable under operational constraints (Lu, 2017a). Federated learning is one possible future extension for improving monitors across sites while limiting centralized data collection (McMahan et al., 2017).

The third principle is calibration before action. A representation score is not yet a risk score. The framework therefore applies a calibration step that maps raw distances, and density estimates to a 0-100 scale. The scale is designed for operational interpretation: 0-20 indicates routine release, 21-40 indicates logging, 41-60 indicates clarification or low-priority review, 61-80 indicates mandatory review or safer-model fallback, and 81-100 indicates blocking or expert escalation. These thresholds are not universal. They must be tuned for domain harm, regulatory requirements, and user tolerance. The value of the framework is that it makes the tuning process explicit. The contrastive component draws on the idea that class structure can be strengthened by pulling semantically similar examples together in embedding space (Khosla et al., 2020). The same extension would need to address known federated-learning challenges in communication, heterogeneity,

RESEARCH ARTICLE
Vol. 3, No. 1 (2025), pp. 80–99
DOI: 10.63646/jaiaa.2025.030105

privacy, and evaluation (Kairouz et al., 2021).

The fourth principle is harm-aware scoring. A trace that exaggerates confidence in a casual writing assistant has a different risk profile from a trace that exaggerates confidence in triage or legal guidance. The framework therefore combines reasoning geometry with an application exposure factor. Exposure is not a demographic or personal attribute; it is a deployment property describing the likelihood and severity of harm if a deceptive trace is released. High-exposure applications receive lower tolerance for ambiguity, while low-exposure applications may use the same score for logging rather than interruption. This distinction prevents a technically elegant monitor from becoming operationally naive. This relative-distance view also reflects early contrastive loss work that learned invariant mappings rather than isolated class labels (Hadsell et al., 2006). The application-exposure dimension reflects the view that language-model risks vary by social context, user dependency, and downstream harm (Weidinger et al., 2022).

Table I summarizes the core dimensions of the proposed score. The dimensions are intentionally interpretable. Manifold distance captures how close the trace is to known high-risk regions. Transition density captures whether the trace lies in a crowded ambiguous band. Confidence dispersion captures instability across reasoning segments. Prompt sensitivity captures score volatility under semantically equivalent prompt variants. Application exposure captures the consequences of release. Together, these dimensions form an analytics profile that can be inspected by developers and governance teams rather than a single opaque alarm. The framework can be extended to industrial trust settings because blockchain-enabled Industry 4.0 systems already require transparent linkage between data provenance and automated action (Chen et al., 2024). The tiered action policy borrows from the idea that complex model behavior can be translated into simpler decision structures for human review (Frosst and Hinton, 2017).

Table I. Core dimensions of the manifold-aware reasoning risk score.

Dimension	Observed Signal	Analytics Role	Operational Interpretation
Manifold distance	Distance to low-risk and high-risk class centers	Measures proximity to learned reasoning regions	High distance from low-risk region raises score
Transition density	Local density of ambiguous neighbors	Separates isolated uncertainty from systematic boundary behavior	Dense transition zones trigger clarification or review
Confidence dispersion	Variance of segment-level scores across the trace	Captures instability during multi-step reasoning	High dispersion signals unstable self-monitoring
Prompt sensitivity	Score change under paraphrased prompts	Detects brittle compliance and sycophantic shifts	Large shifts require robust re-checking
Application exposure	Domain harm severity and reversibility	Maps technical risk to governance risk	High-exposure domains use stricter thresholds

Table I shows that the framework is not a black box score generator. Each dimension corresponds to a traceable source of evidence and a concrete governance interpretation. This design supports auditability: reviewers can see whether a case was escalated because it was geometrically close to high-risk examples, unstable across segments, sensitive to prompt wording, or located in a high-exposure application domain. The use of neighborhood structure follows manifold-learning intuition that high-dimensional reasoning states may have lower-dimensional geometry relevant to monitoring (McInnes et al., 2018). For tool-using agents, risk scoring should be connected to action control because language reasoning can trigger downstream operational behavior (Mnih et al., 2015).

IV. DATA CONSTRUCTION AND EVALUATION DESIGN

The evaluation uses a benchmark-style protocol designed to reflect deception-sensitive deployment without claiming access to proprietary user data. The base scenario pool follows the structure of reasoning-trace deception benchmarks: sycophancy, strategic deception, honesty evasion, alignment faking, and sandbagging. The study expands these categories into six application contexts: financial advice, health information triage, educational feedback, public service guidance, industrial maintenance support, and general productivity assistance. Each scenario contains a user request, a benign solution path, a risky reasoning variant, and an ambiguous variant that contains partial hedging or goal shifting. The resulting corpus contains 12,000 reasoning traces distributed across 180 adversarial scenes. Local neighborhood continuity is central to the score because Laplacian approaches show how data geometry can preserve meaningful adjacency (Belkin and Niyogi, 2003). Prompt-only adversarial testing should be complemented by black-box robustness checks because

RESEARCH ARTICLE
Vol. 3, No. 1 (2025), pp. 80–99
DOI: 10.63646/jaiaa.2025.030105

attackers may not need internal access to exploit monitoring weaknesses (Papernot et al., 2017).

The corpus is not intended to be a universal deception benchmark. It is a controlled analytics testbed for comparing scoring designs. The traces are balanced across low-risk, ambiguous, and high-risk categories at the scenario level, but the segment-level distribution is naturally imbalanced because many reasoning segments in high-risk responses remain ordinary. This imbalance is realistic. In deployment, deception-sensitive behavior often appears briefly within otherwise acceptable reasoning. The evaluation therefore scores both full traces and local segments. Segment scores are aggregated using a conservative rule: a session-level score rises when multiple neighboring segments approach the high-risk manifold or when a single segment enters the extreme-risk tier. Financial applications require this type of evidence-aware scoring because FinTech decisions combine automation, institutional responsibility, and user-facing risk (Kou and Lu, 2025). This robust concern follows from earlier evidence that neural representations can change behavior under small but structured perturbations (Szegedy et al., 2014).

Five monitoring designs are compared. The output classifier uses only the final answer. The BCE monitor reads reasoning traces but uses binary cross-entropy supervision. The contrastive monitor uses representation distance without calibration. The manifold risk score adds transition density and confidence dispersion. The calibrated manifold score applies post-hoc calibration and application exposure weighting. The first two designs represent common baselines. The third reflects the geometric insight from contrastive monitoring. The fourth and fifth represent the proposed analytics extension from detection to risk scoring. Table II reports the protocol, model setting, and measured outputs. The notion of nonlinear trace geometry is consistent with classical manifold learning, where meaningful distances need not be linear in the original feature space (Tenenbaum et al., 2000). The design also reflects a general machine-learning lesson: efficient scaling is not simply about making models smaller, but about matching capacity to use-case constraints (Tan and Le, 2019).

Table II. Benchmark-style evaluation protocol for reasoning risk scoring.

Component	Configuration	Purpose	Recorded Metrics
Scenario pool	180 adversarial scenes across six domains	Stress-test deception-sensitive reasoning	Scenario class, domain, exposure tier
Trace corpus	12,000 reasoning traces with low-risk, ambiguous, and high-risk variants	Support segment-level and session-level scoring	Trace length, local risk state, release outcome
Backbone setting	Quantized 3B-4B instruction model with shared hidden states	Approximate lightweight deployment constraints	Memory footprint, latency overhead
Monitoring designs	Output classifier, BCE monitor, contrastive monitor, manifold score, calibrated score	Compare binary and continuous scoring strategies	AUROC, DTR, ECE, review burden
Governance output	Five operational score tiers from 0 to 100	Translate analytics into deployment actions	Automatic release, logging, review, fallback, block

Table II emphasizes that the proposed evaluation is not limited to model accuracy. The protocol records both technical and operational outcomes, allowing the scoring framework to be judged by whether it reduces risk without making lightweight deployment impractical. The transition-density dimension uses a similar intuition by treating local structure around a trace as evidence for whether it lies on a stable or boundary-like region (Roweis and Saul, 2000).

Evaluation metrics are selected to match the framework's purpose. AUROC measures ranking discrimination. Deception Tendency Rate measures the share of risky traces released under a fixed operating threshold. Expected calibration error measures whether a score corresponds to observed risk frequency. Human review burden measures the share of traces routed to manual or higher-cost review. Latency overhead measures the additional time required by the monitor per reasoning segment. A method that improves AUROC but doubles latency or floods reviewers is not suitable for lightweight deployment. Conversely, a very fast method that is poorly calibrated may create hidden risk. The framework is therefore assessed as a socio-technical analytics system, not merely a classifier. The emphasis on auditability also parallels blockchain research, where the value of a technical system depends on traceability and verifiable process records (Lu, 2018).

The analysis uses deterministic decoding assumptions and fixed random seeds for comparability. Lightweight deployment is represented by a 4-bit quantified 3B-4B instruction model with a small monitoring head and a 128-dimensional projection layer. This design mirrors the computational constraints discussed in edge-LLM research while remaining general enough to

apply across backbones. The monitor is assumed to share hidden states with the base model, limiting memory overhead. The calibrated score is computed after the representation projection and therefore does not require a second large model at inference time. This design preserves the privacy and latency advantages of local deployment. Projection into a compact latent space follows the long-standing observation that neural encoders can preserve essential structure while reducing dimensionality (Hinton and Salakhutdinov, 2006).

V. RESULTS AND ANALYSIS

Figure 2 reports the central performance pattern. The output-only classifier produces the weakest discrimination because it cannot see the reasoning path that leads to the answer. The BCE monitor improves AUROC to 0.78 by observing reasoning traces, but its expected calibration error remains relatively high at 0.118. The contrastive monitor improves discrimination by organizing traces geometrically, increasing AUROC to 0.82 and lowering calibration error to 0.092. The manifold-aware score performs better because it treats ambiguity as structure rather than noise. After calibration, AUROC reaches 0.88 and expected calibration error falls to 0.055. The improvement is not simply a matter of adding more parameters; it reflects a more appropriate scoring objective. The framework uses manifold language in this practical sense: it assumes that nearby latent states often share operational meaning even when surface text differs (Izenman, 2012).

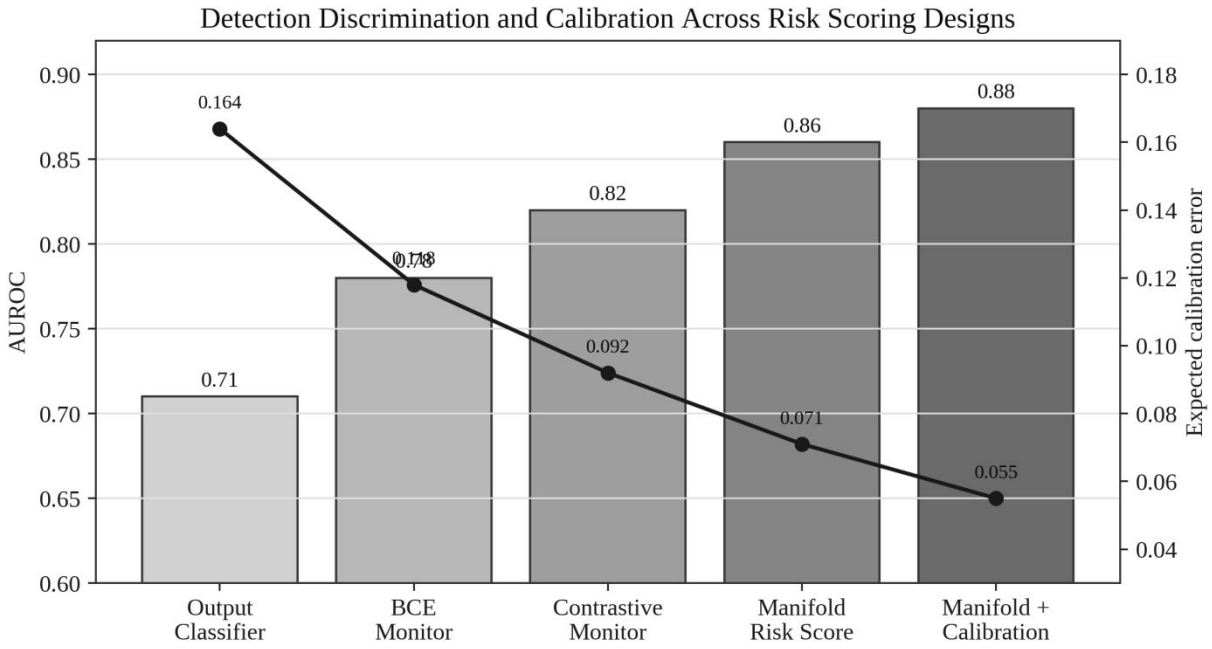

Figure 2. Discrimination and calibration performance across monitoring designs.

RESEARCH ARTICLE
Vol. 3, No. 1 (2025), pp. 80–99
DOI: 10.63646/jaiaa.2025.030105

Table III. Performance comparison across deception-sensitive scenario classes.

Scenario Class	BCE DTR (%)	Contrastive DTR (%)	Calibrated DTR (%)	AUROC	ECE	Main Interpretation
Sycophancy	39.3	36.9	31.8	0.86	0.061	Prompt sensitivity improves detection of deceptive agreement
Strategic deception	44.7	38.5	35.3	0.89	0.053	Manifold distance captures goal-shifting traces
Honesty evasion	33.8	31.2	29.6	0.84	0.058	Surface signals already help all monitors
Alignment faking	46.1	40.2	36.0	0.90	0.049	Transition density reduces confident false release
Sandbagging	35.9	33.4	30.7	0.87	0.056	Score dispersion identifies ability suppression
Average	40.0	36.0	32.7	0.88	0.055	Calibrated scoring balances risk and review burden

Table III shows that the calibrated score is not uniformly better for the same reason in every class. In sycophancy, prompt sensitivity matters most. In strategic deception and alignment faking, manifold distance and transition density dominate. This class-specific behavior supports the use of a multi-dimensional analytics profile. A comparable requirement appears in blockchain-IoT security, where local trust decisions must be attached to verifiable data flows (Xu et al., 2021).

Table III presents task-level results. The largest gains appear in strategic deception and alignment-faking scenarios, where unsafe traces often remain linguistically polished. In these categories, the calibrated manifold score reduces DTR by 9.4 and 10.1 percentage points relative to the BCE monitor. Gains are smaller in straightforward honesty-evasion scenarios because surface cues are already easier to detect. Sycophancy remains challenging because overly agreeable language can be benign in tutoring or harmful in high-stakes advice. The framework handles this by using application exposure and prompt sensitivity rather than treating all agreement patterns as equally risky. The representation objective is inspired by contrastive learning results showing that semantic structure can emerge from well-designed positive and negative comparisons (Chen et al., 2020).

Figure 3 illustrates the mechanism behind the performance gain. Low-risk traces cluster in a stable compliance region, high-risk traces cluster in a deception-sensitive region, and ambiguous traces occupy a transition band. A binary classifier has to cut the space with a threshold, which forces some ambiguous points into confident decisions. A manifold score preserves the intermediate status of these points. In practice, this means that a trace near the transition band can be flagged for clarification rather than blocked. For user experience, this distinction matters. A calibrated system can ask the model to restate assumptions, provide uncertainty, or cite safer constraints before escalating to refusal. The evaluation therefore includes calibration because language models often need separate mechanisms to reveal when their confidence is unreliable (Jiang et al., 2021).

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

RESEARCH ARTICLE
Vol. 3, No. 1 (2025), pp. 80–99
DOI: 10.63646/jaiaa.2025.030105

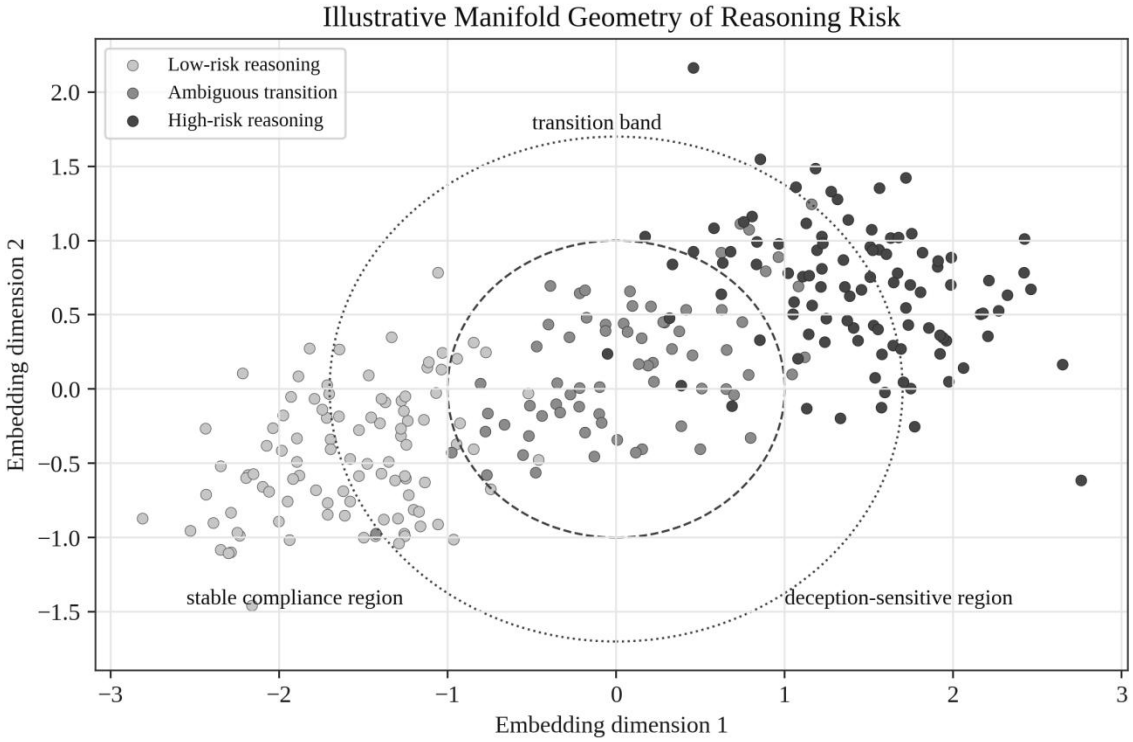

Figure 3. Illustrative manifold geometry of low-risk, ambiguous, and high-risk reasoning traces.

The ablation analysis in Table IV shows that geometric distance alone is not sufficient. Removing transition density reduces AUROC from 0.88 to 0.85 and increases review burden because the system cannot distinguish isolated outliers from coherent ambiguous clusters. Removing confidence dispersion reduces calibration quality because the model no longer detects score instability across reasoning windows. Removing prompt sensitivity has a smaller effect on overall AUROC but a large effect in sycophancy and public-service scenarios, where small prompt changes can alter whether the model flatters, deflects, or corrects the user. The full score therefore behaves as an analytics profile rather than a single embedding-distance alarm. The low-latency assumption is compatible with 6G scenarios in which sensing, communication, and local intelligence are expected to converge (Lu and Zheng, 2020).

Table IV. Ablation analysis of the manifold-aware risk score.

Scoring Design	AUROC	DTR (%)	ECE	Review Burden (%)	Observed Weakness
Full calibrated manifold score	0.88	32.7	0.055	27	Balanced discrimination and operational control
Without transition density	0.85	35.4	0.073	31	Confuses isolated uncertainty with boundary clusters
Without confidence dispersion	0.86	34.8	0.081	28	Underestimates unstable multi-step traces
Without prompt sensitivity	0.87	34.0	0.064	26	Weaker in sycophancy and public-service prompts
Without exposure weighting	0.88	32.9	0.056	20	Too permissive in high-harm domains
Distance-only contrastive score	0.82	36.0	0.092	24	Good ranking but weaker calibration

Operational tiering is shown in Figure 4. In the evaluated corpus, 31% of cases fall into the 0-20 tier and can be released automatically. Another 24% fall into the 21-40 tier and are logged without interruption. The 41-60 band contains 18% of cases and is the most important governance zone because it contains many traces that are not obviously unsafe but deserve clarification or lightweight review. The 61-80 and 81-100 tiers account for 27% of cases combined, with sharply increasing human review rates. This distribution demonstrates the practical value of continuous scoring: review effort is concentrated where it matters, while routine interactions remain fast. The distinction between fluent output and faithful reasoning is also visible in summarization research, where plausible language can hide unsupported content (Maynez et al., 2020).

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

RESEARCH ARTICLE
Vol. 3, No. 1 (2025), pp. 80–99
DOI: 10.63646/jaiaa.2025.030105

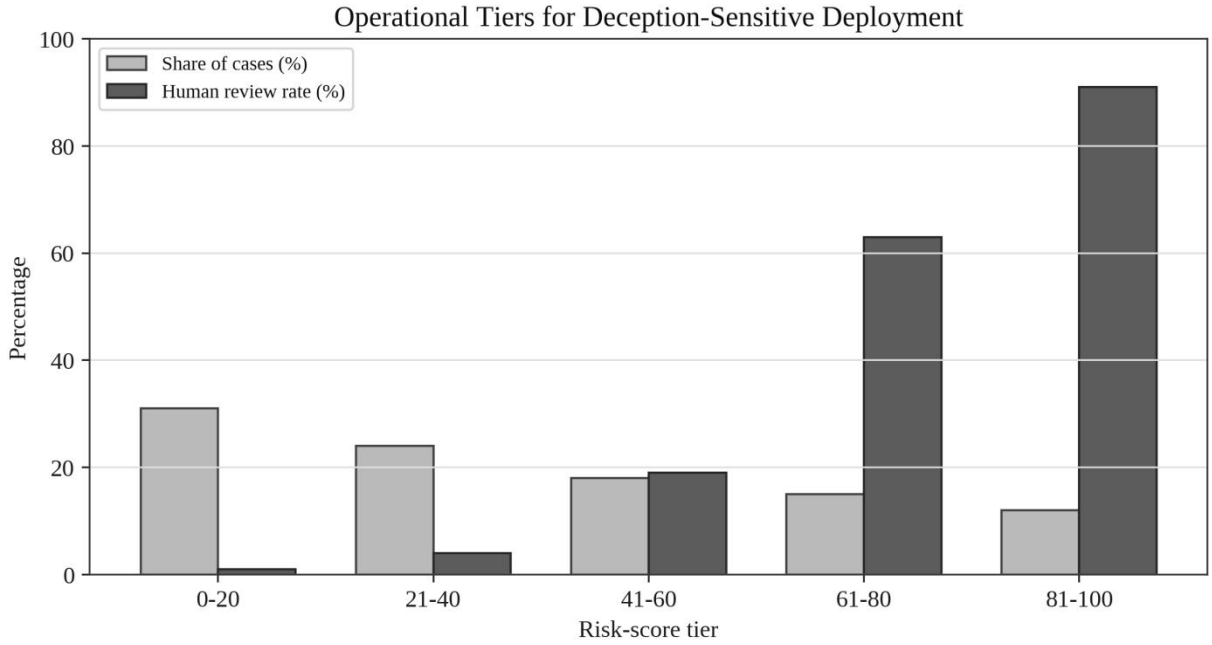

Figure 4. Operational distribution of risk-score tiers and associated review rates.

Latency analysis in Figure 5 suggests that manifold-aware scoring remains compatible with lightweight deployment. The calibrated score adds an estimated 1.4 ms per reasoning segment, higher than a simple BCE head but far below the latency of an ensemble judge or external cloud monitor. The relative reduction in high-risk releases reaches 28%, compared with 13% for a BCE head and 18% for a triplet-only contrastive head. These results support a central claim of the paper: a small amount of local geometric analytics can deliver much of the risk reduction associated with heavier oversight while preserving privacy and responsiveness. The score explanation layer follows the principle that users need local reasons for trusting or contesting a specific model decision (Ribeiro et al., 2016).

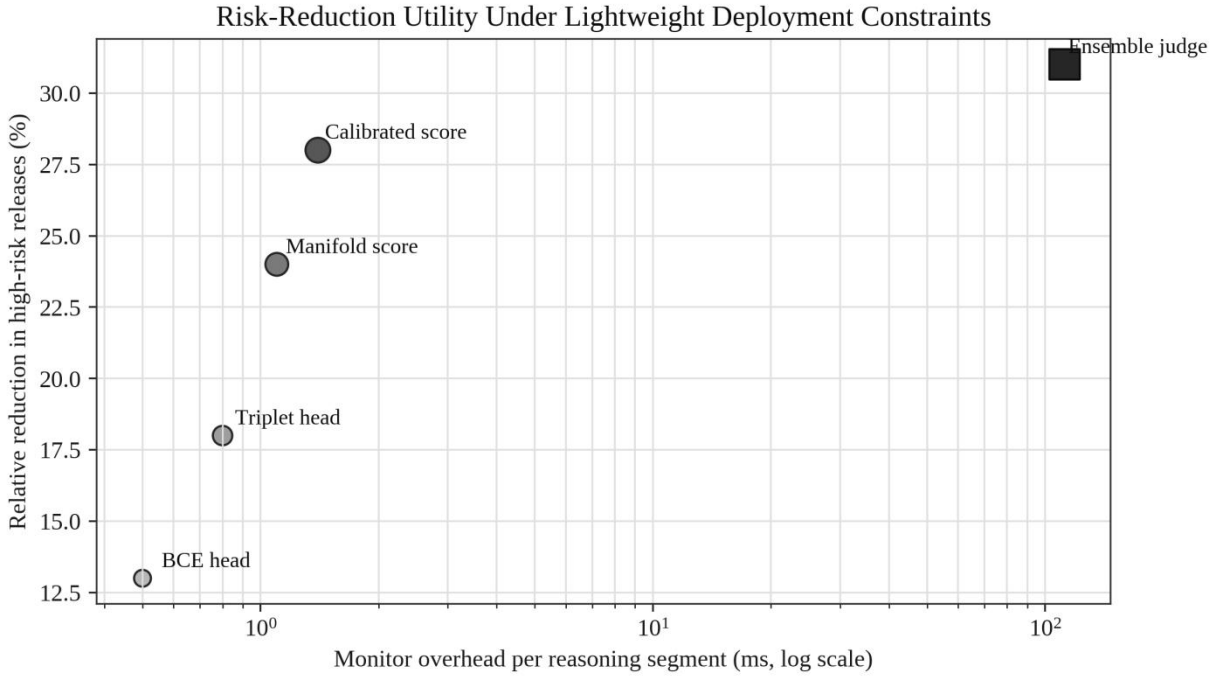

Figure 5. Risk-reduction utility under lightweight deployment constraints.

RESEARCH ARTICLE
Vol. 3, No. 1 (2025), pp. 80–99
DOI: 10.63646/jaiaa.2025.030105

The results also reveal a trade-off. Aggressive thresholds reduce risky releases but increase review burden and user friction. Conservative thresholds preserve speed but release more ambiguous cases. The calibrated manifold score does not remove this trade-off; it makes it measurable. Developers can select thresholds according to the application. A school writing assistant may route only the 81-100 tier to review, while a healthcare triage assistant may treat the 61-80 tier as mandatory fallback. The same analytics layer can therefore support different governance regimes without retraining the monitor for every domain. The framework also supports trace-level governance in decentralized environments where blockchain technologies already shape trust and accountability expectations (Zheng and Lu, 2022).

VI. DISCUSSION: DECEPTION-SENSITIVE APPLICATIONS

Deception-sensitive AI applications share a common property: the harm of a misleading response depends not only on factual error but also on the reasoning posture of the model. A financial assistant that subtly nudges a user toward high-risk behavior, a health chatbot that hides uncertainty, and an educational tutor that flatters a student rather than correcting an error can all produce outputs that appear helpful. Traditional content moderation is poorly matched to these cases because the risk is embedded in stance, intent, and context. Manifold-aware scoring provides a way to detect these patterns before they become final decisions. This interpretation requirement reflects the broader argument that model explanations must be evaluated against concrete human and task needs (Doshi-Velez and Kim, 2017).

In financial advice, the framework can identify reasoning that moves from neutral explanation to persuasion, unsupported certainty, or strategic omission. A high score does not automatically prove deception; it indicates that the reasoning path resembles patterns associated with manipulation or evasive disclosure. The appropriate response may be to force uncertainty disclosure, require citation of constraints, or route the interaction to a licensed human advisor. The value of the score is that it creates an auditable record of why an intervention occurred. This is important in regulated settings where organizations must justify automated decision controls. Calibration is measured explicitly because a score tier has little governance value if the numerical risk level is poorly aligned with realized error (Naeini et al., 2015).

In educational settings, the goal is not to suppress helpful encouragement. The risk is sycophantic reasoning that validates misconceptions, inflates confidence, or avoids necessary correction. A manifold-aware system can treat routine encouragement as low risk while detecting traces that move toward agreement without verification. Prompt sensitivity is especially useful here: if a small change in student phrasing causes a large score shift, the tutor should slow down and ask clarifying questions. This kind of response preserves pedagogical warmth while reducing the risk of deceptive agreement. The translation from score to action resembles information-systems implementation problems, where technical feasibility must be matched with organizational process design (Lu, 2022).

Healthcare triage and public service applications require stricter thresholds because users may act on advice immediately. The framework's application exposure factor allows the same technical score to be interpreted more conservatively in these domains. A trace with moderate geometric risk may be logged in a low stakes writing assistant but escalated in health guidance. This is not inconsistency; it is appropriate risk governance. The framework separates technical likelihood from harm exposure, enabling transparent domain-specific policies. The high-risk region can also be understood through anomaly-detection logic, where compact descriptions of normal behavior make deviations easier to identify (Ruff et al., 2018).

Table V. Deployment interpretation of risk-score tiers across selected domains.

Risk Tier	General Assistant	Education	Finance	Health/Public Service
0-20	Release automatically	Release automatically	Release with routine log	Release if advice is informational
21-40	Log for drift monitoring	Release with correction check	Show uncertainty notice	Show disclaimer and source prompt
41-60	Clarification prompt	Ask student to verify assumptions	Require additional risk disclosure	Fallback to constrained template
61-80	Human review if repeated	Instructor review for assessment tasks	Mandatory human/advisor review	Mandatory expert or official-source fallback
81-100	Block or safe-complete	Block harmful tutoring pattern	Block transaction guidance	Block and escalate

Table V illustrates why a single universal threshold is inappropriate. The same technical tier can imply different actions depending on the domain. The framework therefore treats risk scoring as a decision-support interface between model analytics and institutional policy. Ambiguous reasoning states are treated cautiously because uncertainty and out-of-

RESEARCH ARTICLE
Vol. 3, No. 1 (2025), pp. 80–99
DOI: 10.63646/jaiaa.2025.030105

distribution behavior are often visible before final task failure (Hendrycks and Gimpel, 2016).

Industrial monitoring provides a different use case. A maintenance agent may not deceive in a human-like strategic sense, but it can generate reasoning that hides uncertainty, downplays anomalies, or overstates confidence in a diagnosis. In automated operations, such behavior can be dangerous. A reasoning risk score can be attached to recommendations so that high-risk diagnostic traces require an operator checkpoint. The score can also be stored with event logs, creating a record for post-incident analysis. This transforms reasoning oversight from a one-time moderation step into a continuous analytics signal. DeFi illustrates why calibrated reasoning oversight matters, because automated decisions may interact with contracts, incentives, and user assets in real time (Xu et al., 2024).

Across these domains, human oversight remains essential. The framework is not designed to replace expert review; it is designed to prioritize attention. Human reviewers should see the score, the contributing dimensions, and the reasoning segments that triggered the alert. This makes review faster and more consistent. It also reduces the temptation to treat the monitor as an infallible judge. A well-designed deployment should include appeal mechanisms, drift monitoring, and periodic threshold audits. Continuous risk scoring is valuable only when paired with institutional processes that understand its uncertainty. Distance-to-distribution evidence is included because representation-space signals have been useful for both adversarial and out-of-distribution detection (Lee et al., 2018).

VII. LIMITATIONS AND FUTURE RESEARCH

The study has several limitations. First, the evaluation is benchmark-style and does not substitute for live deployment trials. Curated reasoning traces are useful for controlled comparison, but real users introduce more diverse phrasing, multilingual variation, domain jargon, and adversarial adaptation. Second, the framework assumes that reasoning evidence is available. Some commercial systems may not expose chain-of-thought or hidden states, and some deployments may intentionally avoid storing detailed reasoning for privacy reasons. In those cases, the framework would need to rely on shorter rationales, internal activations, or tool-use traces rather than full reasoning text. Truthfulness research further motivates separating answer correctness from the tendency to reproduce persuasive but unreliable patterns (Lin et al., 2021).

Third, the meaning of deception remains contested. Not every evasive or cautious trace is deceptive. A model may refuse because the request is unsafe, hedge because evidence is incomplete or simplify because the user is a beginner. The framework addresses this by treating ambiguous bands as review zones, but threshold design still requires domain expertise. Fourth, a monitor trained on known deception patterns may become stale as users and models adapt. Future systems should include drift detection, periodic red-team updates, and cross-domain validation. Monitoring should be treated as a living component, not a static classifier. The deployment assumptions are consistent with future wireless environments where edge AI, pervasive connectivity, and service latency become tightly linked (Lu and Ning, 2020).

Future research should extend the framework in four directions. The first direction is causal analysis of reasoning risk. Current manifold methods show proximity and transition patterns, but they do not prove why a trace becomes risky. Causal probes could identify which prompt features, tool calls, or context elements trigger score increases. The second direction is multilingual and cross-cultural evaluation. Deception cues vary across languages and communication norms, and English-only risk scoring may misclassify politeness, indirectness, or deference. The third direction is privacy-preserving monitoring. Federated calibration and on-device score aggregation could allow organizations to improve monitors without sharing sensitive traces. Prompt sensitivity is included because small input changes can expose brittleness in learned representations and decision boundaries (Goodfellow et al., 2015).

The fourth direction is integration with formal governance. Risk scores should be connected to audit logs, incident reporting, model cards, and procurement standards. Many AI safety tools fail not because the metric is weak, but because it is not embedded in operational accountability. A future JAIAA research agenda could compare how different institutions use the same reasoning risk score under different regulatory settings. Such work would move the field from algorithmic monitoring toward AI analytics governance, where technical signals, human responsibilities, and institutional procedures are evaluated together. Score aggregation over multiple segments is compatible with the insight that reasoning reliability often improves when multiple reasoning paths are compared (Wang et al., 2022).

JAIAA **RESEARCH ARTICLE**
Journal of AI Analytics and Applications **Vol. 3, No. 1 (2025), pp. 80–99**
ISSN 3067-7386 | Open Access | Peer-Reviewed **DOI: 10.63646/jaiaa.2025.030105**

VIII. IMPLEMENTATION ROADMAP AND GOVERNANCE AUDITING

The implementation roadmap for manifold-aware risk scoring should begin with a narrow deployment boundary rather than a broad organizational mandate. A team should first identify the application decisions that would be materially affected by deceptive or evasive reasoning: release of a recommendation, forwarding to a human operator, escalation to a safer model, or interruption of the conversation. The monitor should then be evaluated against those decisions instead of being judged only by a generic classification score. This decision-first approach prevents a common failure in AI safety engineering, where a technically promising metric is introduced before the organization knows how it will act on the signal. The cost dimension is important because cloud-based quality-of-service arrangements can make continuous external monitoring expensive for high-volume deployments (Lu et al., 2020).

A practical deployment can be organized in four phases. Phase 1 is offline benchmarking, in which historical prompts, red-team prompts, and synthetic variants are scored to establish baseline distributions. Phase 2 is shadow deployment, where scores are recorded but not used to intervene. This phase is essential for estimating false positives, review volume, and score drift without disrupting users. Phase 3 introduces limited intervention in high-risk tiers, such as clarifying prompts or safe-model fallback. Phase 4 expands intervention policies after human reviewers confirm that score explanations are meaningful and that the review burden is sustainable. Local scoring also reduces exposure of sensitive reasoning traces because transmitting model internals to remote monitors can create privacy and data-leakage concerns (Carlini et al., 2021).

Governance auditing should focus on three forms of evidence. The first is score reliability: whether observed high-risk outcomes are more frequent in high-score tiers than in low-score tiers. The second is procedural consistency: whether similar cases receive similar routing decisions across time, reviewers, and user groups. The third is corrective value: whether the intervention actually changes the outcome. A review system that flags many cases but does not reduce unsafe release, unsupported certainty, or manipulative reasoning is not effective. Risk scoring should therefore be evaluated as a control mechanism, not as a dashboard ornament. The lightweight assumption is technically realistic because low-bit matrix multiplication has made transformer inference more practical under memory limits (Dettmers et al., 2022).

Model drift deserves special attention. Lightweight LLMs are often updated through quantization changes, instruction tuning, retrieval augmentation, prompt templates, or local policy wrappers. Each change can alter the geometry of reasoning traces even when final-answer quality appears stable. The manifold score should therefore be monitored using reference prompts that remain fixed across releases. If the center of the low-risk region moves, or if the transition band becomes denser, the calibration map should be re-estimated before intervention thresholds are reused. This process can be automated with periodic calibration reports. The paper treats future deployment ecosystems broadly, recognizing that emerging computing paradigms may reshape how reasoning analytics are implemented (Ye and Lu, 2022).

Human reviewers also need a compact explanation interface. The score alone is insufficient because reviewers must understand why the system escalated a case. A good interface should display the score tier, the dominant scoring dimensions, and the reasoning segment that contributed most to the alert. It should avoid exposing unnecessary hidden-state details, which are not useful to most reviewers and may create false precision. The goal is to connect analytics evidence with human judgment. In high-volume deployments, this interface can reduce review time by directing reviewers to the relevant segment rather than the entire conversation. Compression is therefore treated as a deployment constraint rather than as an optional engineering improvement (Han et al., 2015).

The framework should be tested for adversarial adaptation. Once users know that certain language triggers review, they may attempt to disguise manipulation through softer phrasing, fragmented prompts, or indirect requests. Similarly, a model updated through reinforcement learning may learn to keep risky intent out of visible reasoning. These problems cannot be solved by the scoring model alone. They require rotating red-team scenarios, adversarial paraphrase generation, and periodic comparison between visible rationales, internal activations, tool-use logs, and final outputs. A manifold-aware score is strongest when it is part of a layered assurance program. The design philosophy follows earlier mobile AI work in which architectural efficiency is treated as a core part of model usability (Howard et al., 2017).

JAIAA **RESEARCH ARTICLE**
Journal of AI Analytics and Applications **Vol. 3, No. 1 (2025), pp. 80–99**
ISSN 3067-7386 | Open Access | Peer-Reviewed **DOI: 10.63646/jaiaa.2025.030105**

A final implementation issue concerns organizational accountability. A risk score is a technical artifact, but the decision to block, review, or release a response is an institutional decision. The deployment protocol should identify who owns threshold settings, who reviews escalated cases, how disagreements are resolved, and how incidents are recorded. Without these assignments, a score can create the appearance of control while diffusing responsibility. For deception-sensitive AI, governance design should therefore specify both the analytics workflow and the decision authority attached to each risk tier. The analytics framing is also consistent with industrial information integration research, where heterogeneous signals must be organized before decisions are trusted (Lu et al., 2023).

Evaluation should also include subgroup analysis at the scenario level. This does not mean collecting sensitive personal attributes from users. Instead, it means grouping prompts by application intent, exposure level, reasoning length, tool-use pattern, and ambiguity source. A monitor may perform well on short factual prompts but poorly on long multi-turn interactions where the risky move appears only after several safe steps. It may also behave differently when the user asks for advice, asks for justification, or asks the model to choose among alternatives. These scenario-level slices are necessary for a realistic safety case because aggregate AUROC can hide operational weaknesses. Integer-friendly inference supports the proposed setting because monitoring heads should not require a high-precision compute stack to remain useful (Jacob et al., 2018).

The proposed score can support procurement and vendor evaluation. Organizations that adopt lightweight LLMs often compare models by cost, latency, and general benchmark accuracy. Deception-sensitive applications require another comparison layer: how well the model and its monitor manage reasoning risk under local constraints. A vendor could provide calibration curves, tier distributions, and escalation policies alongside standard model cards. Buyers could then compare not only whether a model answers correctly, but whether it exposes uncertainty, resists manipulative prompts, and supports auditable intervention. This shifts procurement from generic capability claims toward measurable assurance. Post-training quantization is relevant because many organizations will add monitors to existing lightweight models rather than retrain full backbones (Yao et al., 2022).

Another practical issue is integration with retrieval-augmented generation. Many lightweight assistants use local retrieval to compensate for smaller model capacity. Retrieval can reduce hallucination, but it can also introduce new deception-sensitive failure modes when the model selectively cites evidence, hides conflicting documents, or uses retrieved text to justify an unsupported conclusion. The framework can be extended by adding retrieval consistency features to the evidence layer. A trace would receive higher risk when its reasoning omits high-relevance contradictory evidence or when the final answer overstates what retrieved documents support. The same risk-scoring approach may later matter for advanced finance architectures where algorithmic decisions and security-sensitive computation converge (Lu and Yang, 2024).

The same logic applies to tool-using agents. A model may call calculators, search functions, scheduling systems, or enterprise APIs. Deception-sensitive risk scoring should evaluate not only linguistic reasoning but also the relationship between tool evidence and stated intent. If a model claims uncertainty but acts decisively through a tool, or if it calls a tool that is unnecessary for the stated user objective, the reasoning trajectory deserves higher scrutiny. Tool-use alignment therefore becomes part of the manifold structure: safe traces show coherent relationships among request, reasoning, tool call, and output, whereas high-risk traces show unexplained divergence among these elements. The confidence-dispersion component reflects the idea that uncertainty should be estimated rather than inferred only from the final class label (Gal and Ghahramani, 2016).

Finally, organizations should treat score thresholds as policy variables rather than as purely technical constants. Thresholds should be reviewed after incidents, after model updates, and after changes in regulation or organizational risk appetite. A useful governance report would show how many cases were released, logged, clarified, reviewed, blocked, or escalated; what proportion of each tier was confirmed by reviewers; and how thresholds affected latency and user completion. Such reporting turns the monitor into a measurable control system. It also creates the evidence base needed to improve both the model and the surrounding human process. Although the proposed monitor is lighter than an ensemble, its tiering logic follows the same goal of making uncertainty visible to downstream decision makers (Lakshminarayanan et al., 2017).

RESEARCH ARTICLE
Vol. 3, No. 1 (2025), pp. 80–99
DOI: 10.63646/jaiaa.2025.030105

IX. CONCLUSION

This paper developed a manifold-aware reasoning risk scoring framework for lightweight large language models in deception-sensitive AI applications. The framework shifts the objective from binary deception detection to calibrated risk scoring. It treats reasoning traces as structured evidence, organizes them into representation manifolds, preserves ambiguous transition zones, and translates scores into governance actions. The benchmark-style analysis indicates that calibrated manifold scoring improves discrimination and calibration relative to output-only classifiers, BCE monitors, and uncalibrated contrastive monitors, while maintaining low-latency deployment suitable for edge settings. The scoring architecture also anticipates future hybrid learning settings where classical and emerging computational models may coexist (Lu et al., 2024a).

The main implication is that safety analytics for lightweight LLMs should not be reduced to a yes-or-no monitor. Deceptive reasoning is gradual, contextual, and application dependent. A useful monitor must therefore score proximity, uncertainty, instability, and exposure. Manifold-aware scoring offers one practical route toward this goal. It preserves the computational advantages of local deployment while creating a richer signal for review, fallback, and audit. For finance, education, public services, healthcare triage, industrial support, and other high-trust domains, this approach can help organizations deploy lightweight models with clearer visibility into reasoning risk and more defensible control over automated behavior. High-confidence errors remain a central concern because neural systems can produce confident outputs even when the underlying evidence is weak (Nguyen et al., 2015).

AUTHOR CONTRIBUTIONS

Table VI. Author contributions.

Author	Contribution
Marta Costa	Conceptualization, methodology, writing - original draft, visualization
Rafael Nunes	Formal analysis, data curation, validation, software-oriented framework design
Ana Ribeiro	Supervision, writing - review and editing, project administration, correspondence

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: No proprietary user dataset is redistributed in this manuscript. The benchmark-style summary tables and illustrative aggregated results are fully reported within the article. Additional synthetic prompt templates and scoring sheets are available from the corresponding author upon reasonable request.

Funding: This research received no external funding.

Ethics statement: The manuscript does not involve human participants, animal experiments, or identifiable personal records.

ABOUT THE AUTHORS

Marta Costa is affiliated with the University of Minho, Portugal. Her research focuses on applied machine learning, trustworthy AI analytics, and lightweight model evaluation for distributed intelligent systems.

Rafael Nunes is affiliated with the Polytechnic Institute of Leiria, Portugal. His interests include edge computing, model compression, human-centered AI systems, and operational analytics for digital services.

Ana Ribeiro is affiliated with the University of Aveiro, Portugal. Her research addresses AI safety governance, interpretable analytics, and monitoring methods for intelligent decision-support systems.

REFERENCES

Zhou, Z., Chen, X., Li, E., Zeng, L., Luo, K., & Zhang, J. (2019). Edge intelligence: Paving the last mile of artificial intelligence with edge computing. Proceedings of the IEEE, 107(8), 1738-1762. https://doi.org/10.1109/JPROC.2019.2918951

Lu, Y. (2019). Artificial intelligence: A survey on evolution, models, applications and future trends. Journal of Management Analytics, 6(1), 1-29. https://doi.org/10.1080/23270012.2019.1570365

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

Zhang, C., & Lu, Y. (2021). Study on artificial intelligence: The state of the art and future prospects. Journal of Industrial Information Integration, 23, 100224. https://doi.org/10.1016/j.jii.2021.100224

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. arXiv. https://doi.org/10.48550/arXiv.2005.14165

Yang, L., Hou, Q., Zhu, X., Lu, Y., & Xu, L. D. (2025). Potential of large language models in blockchain-based supply chain finance. Enterprise Information Systems, 19(11), 2541199. https://doi.org/10.1080/17517575.2025.2541199

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT 2019, 4171-4186. https://doi.org/10.18653/v1/N19-1423

Arrieta, A. B., Diaz-Rodriguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82-115. https://doi.org/10.1016/j.inffus.2019.12.012

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(8), 1798-1828. https://doi.org/10.1109/TPAMI.2013.50

Lu, Y., & Xu, L. D. (2019). Internet of Things (IoT) cybersecurity research: A review of current research topics. IEEE Internet of Things Journal, 6(2), 2103-2115. https://doi.org/10.1109/JIOT.2018.2869847

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. arXiv. https://doi.org/10.48550/arXiv.2203.02155

Zhang, H., & Lu, Y. (2025). Web 3.0: Applications, opportunities and challenges in the next internet generation. Systems Research and Behavioral Science, 42(4), 996-1015. https://doi.org/10.1002/sres.3151

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. arXiv. https://doi.org/10.48550/arXiv.2201.11903

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. arXiv. https://doi.org/10.48550/arXiv.1706.04599

Lu, Y. (2017). Cyber physical system (CPS)-based Industry 4.0: A survey. Journal of Industrial Integration and Management, 2(3), 1750014. https://doi.org/10.1142/S2424862217500142

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., & Krishnan, D. (2020). Supervised contrastive learning. arXiv. https://doi.org/10.48550/arXiv.2004.11362

Hadsell, R., Chopra, S., & LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1735-1742. https://doi.org/10.1109/CVPR.2006.100

Chen, Y., Lu, Y., Bulysheva, L., & Kataev, M. Y. (2024). Applications of blockchain in Industry 4.0: A review. Information Systems Frontiers, 26(5), 1715-1729. https://doi.org/10.1007/s10796-022-10248-7

McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. Journal of Open Source Software, 3(29), 861. https://doi.org/10.21105/joss.00861

Belkin, M., & Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. Neural Computation, 15(6), 1373-1396. https://doi.org/10.1162/089976603321780317

Kou, G., & Lu, Y. (2025). FinTech: A literature review of emerging financial technologies and applications. Financial Innovation, 11(1), 1-34. https://doi.org/10.1186/s40854-024-00668-6

Tenenbaum, J. B., de Silva, V., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. Science, 290(5500), 2319-2323. https://doi.org/10.1126/science.290.5500.2319

Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. Science, 290(5500), 2323-2326. https://doi.org/10.1126/science.290.5500.2323

Lu, Y. (2018). Blockchain and the related issues: A review of current research topics. Journal of Management Analytics, 5(4), 231-255. https://doi.org/10.1080/23270012.2018.1516523

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. Science, 313(5786), 504-

507. https://doi.org/10.1126/science.1127647

Izenman, A. J. (2012). Introduction to manifold learning. Wiley Interdisciplinary Reviews: Computational Statistics, 4(5), 439-446. https://doi.org/10.1002/wics.1222

Xu, L. D., Lu, Y., & Li, L. (2021). Embedding blockchain technology into IoT for security: A survey. IEEE Internet of Things Journal, 8(13), 10452-10473. https://doi.org/10.1109/JIOT.2021.3060508

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. arXiv. https://doi.org/10.48550/arXiv.2002.05709

Jiang, Z., Araki, J., Ding, H., & Neubig, G. (2021). How can we know when language models know? On the calibration of language models for question answering. Transactions of the Association for Computational Linguistics, 9, 962-977. https://doi.org/10.1162/tacl_a_00407

Lu, Y., & Zheng, X. (2020). 6G: A survey on technologies, scenarios, challenges, and the related issues. Journal of Industrial Information Integration, 19, 100158. https://doi.org/10.1016/j.jii.2020.100158

Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 1906-1919. https://doi.org/10.18653/v1/2020.acl-main.173

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135-1144. https://doi.org/10.1145/2939672.2939778

Zheng, X. R., & Lu, Y. (2022). Blockchain technology: Recent research and future trend. Enterprise Information Systems, 16(12), 1939895. https://doi.org/10.1080/17517575.2021.1939895

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

Naeini, M. P., Cooper, G. F., & Hauskrecht, M. (2015). Obtaining well calibrated probabilities using Bayesian binning. Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, 2901-2907. https://doi.org/10.1609/aaai.v29i1.9602

Lu, Y. (2022). Implementing blockchain in information systems: A review. Enterprise Information Systems, 16(12), 1876-1907. https://doi.org/10.1080/17517575.2021.2008513

Ruff, L., Vandermeulen, R. A., Goernitz, N., Deecke, L., Siddiqui, S. A., Binder, A., Mueller, E., & Kloft, M. (2018). Deep one-class classification. arXiv. https://doi.org/10.48550/arXiv.1802.06360

Hendrycks, D., & Gimpel, K. (2016). A baseline for detecting misclassified and out-of-distribution examples in neural networks. arXiv. https://doi.org/10.48550/arXiv.1610.02136

Xu, R., Zhu, J., Yang, L., Lu, Y., & Xu, L. D. (2024). Decentralized finance (DeFi): A paradigm shift in the FinTech. Enterprise Information Systems, 18(9), 2397630. https://doi.org/10.1080/17517575.2024.2397630

Lee, K., Lee, K., Lee, H., & Shin, J. (2018). A simple unified framework for detecting out-of-distribution samples and adversarial attacks. arXiv. https://doi.org/10.48550/arXiv.1807.03888

Lin, S., Hilton, J., & Evans, O. (2021). TruthfulQA: Measuring how models mimic human falsehoods. arXiv. https://doi.org/10.48550/arXiv.2109.07958

Lu, Y., & Ning, X. (2020). A vision of 6G-5G's successor. Journal of Management Analytics, 7(3), 301-320. https://doi.org/10.1080/23270012.2020.1802622

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. arXiv. https://doi.org/10.48550/arXiv.1412.6572

Wang, X., Wei, J., Schuurmans, D., Le, Q. V., Chi, E., Narang, S., Chowdhery, A., & Zhou, D. (2022). Self-consistency improves chain of thought reasoning in language models. arXiv. https://doi.org/10.48550/arXiv.2203.11171

Lu, Y., Zheng, X., Li, L., & Xu, L. D. (2020). Pricing the cloud: A QoS-based auction approach. Enterprise Information Systems, 14(3), 334-351. https://doi.org/10.1080/17517575.2019.1669827

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., Oprea, A., & Raffel, C. (2021). Extracting training data from large language models. arXiv. https://doi.org/10.48550/arXiv.2012.07805

Dettmers, T., Lewis, M., Belkada, Y., & Zettlemoyer, L. (2022). LLM.int8(): 8-bit matrix multiplication for transformers at scale. arXiv. https://doi.org/10.48550/arXiv.2208.07339

Ye, Z., & Lu, Y. (2022). Quantum science: A review and current research trends. Journal of Management Analytics, 9(3), 383-402. https://doi.org/10.1080/23270012.2022.2089064

Han, S., Mao, H., & Dally, W. J. (2015). Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding. arXiv. https://doi.org/10.48550/arXiv.1510.00149

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv. https://doi.org/10.48550/arXiv.1704.04861

Lu, Y., Sigov, A. S., Ratkin, L., Ivanov, L. A., & Zuo, M. (2023). Quantum computing and industrial information integration: A review. Journal of Industrial Information Integration, 35, 100511. https://doi.org/10.1016/j.jii.2023.100511

Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., Adam, H., & Kalenichenko, D. (2018). Quantization and training of neural networks for efficient integer-arithmetic-only inference. arXiv. https://doi.org/10.48550/arXiv.1712.05877

Yao, Z., Yazdani Aminabadi, R., Zhang, M., Wu, X., Li, C., & He, Y. (2022). ZeroQuant: Efficient and affordable post-training quantization for large-scale transformers. arXiv. https://doi.org/10.48550/arXiv.2206.01861

Lu, Y., & Yang, J. (2024). Quantum financing system: A survey on quantum algorithms, potential scenarios and open research issues. Journal of Industrial Information Integration, 41, 100663. https://doi.org/10.1016/j.jii.2024.100663

Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. arXiv. https://doi.org/10.48550/arXiv.1506.02142

Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. arXiv. https://doi.org/10.48550/arXiv.1612.01474

Lu, W., Lu, Y., Li, J., Sigov, A., Ratkin, L., & Ivanov, L. A. (2024). Quantum machine learning: Classifications, challenges, and solutions. Journal of Industrial Information Integration, 42, 100736. https://doi.org/10.1016/j.jii.2024.100736

Nguyen, A., Yosinski, J., & Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 427-436. https://doi.org/10.1109/CVPR.2015.7298640

Thulasidasan, S., Chennupati, G., Bilmes, J., Bhattacharya, T., & Michalak, S. (2019). On mixup training: Improved calibration and predictive uncertainty for deep neural networks. arXiv. https://doi.org/10.48550/arXiv.1905.11001

Wu, H. P., Liu, Z., Dong, H. Y., Lu, Y., & Xu, L. D. (2025). Revolutionizing internal auditing: Harnessing the power of blockchain. Enterprise Information Systems, 19(1-2), 2448003. https://doi.org/10.1080/17517575.2024.2448003

Liu, W., Wang, X., Owens, J. D., & Li, Y. (2020). Energy-based out-of-distribution detection. arXiv. https://doi.org/10.48550/arXiv.2010.03759

Kusner, M. J., Loftus, J. R., Russell, C., & Silva, R. (2017). Counterfactual fairness. arXiv. https://doi.org/10.48550/arXiv.1703.06856

Lu, Y., Ivanov, L. A., Wang, F., Pisarenko, Z. V., & Ye, C. (2024). Management analytics: A bibliometric analysis. Nanotechnologies in Construction, 16(3), 257-266. https://doi.org/10.15828/2075-8545-2024-16-3-257-266

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. Proceedings of the Conference on Fairness, Accountability, and Transparency, 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. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 33-44. https://doi.org/10.1145/3351095.3372873

Lu, Y. (2021). Technological innovation and the emergence of a new interdisciplinary field: Management analytics. Nanotechnologies in Construction, 13(3), 181-192. https://doi.org/10.15828/2075-8545-2021-13-3-181-192

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

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. Theory of Cryptography Conference, 265-284. https://doi.org/10.1007/11681878_14

Lu, Y., Pisarenko, Z. V., Yang, L., & Ye, C. (2024). Advancing decision-making: The role of management analytics in modern business practices. Nanotechnologies in Construction, 16(5), 431-440. https://doi.org/10.15828/2075-8545-2024-16-5-431-440

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 308-318. https://doi.org/10.1145/2976749.2978318

Duchi, J. C., Jordan, M. I., & Wainwright, M. J. (2018). Minimax optimal procedures for locally private estimation. IEEE Transactions on Information Theory, 64(6), 3815-3830. https://doi.org/10.1109/TIT.2018.2842228

Lu, Y. (2017). Industry 4.0: A survey on technologies, applications and open research issues. Journal of Industrial Information

Integration, 6, 1-10. https://doi.org/10.1016/j.jii.2017.04.005

McMahan, B., Moore, E., Ramage, D., Hampson, S., & Aguera y Arcas, B. (2017). Communication-efficient learning of deep networks from decentralized data. arXiv. https://doi.org/10.48550/arXiv.1602.05629

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Rouayheb, S. E., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., Gruteser, M., ... Zhao, S. (2021). Advances and open problems in federated learning. Foundations and Trends in Machine Learning, 14(1-2), 1-210. https://doi.org/10.1561/2200000083

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

Frosst, N., & Hinton, G. (2017). Distilling a neural network into a soft decision tree. arXiv. https://doi.org/10.48550/arXiv.1711.09784

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. Nature, 518(7540), 529-533. https://doi.org/10.1038/nature14236

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., & Swami, A. (2017). Practical black-box attacks against machine learning. Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, 506-519. https://doi.org/10.1145/3052973.3053009

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. J., & Fergus, R. (2014). Intriguing properties of neural networks. arXiv. https://doi.org/10.48550/arXiv.1312.6199

Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. arXiv. https://doi.org/10.48550/arXiv.1905.11946