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

RESEARCH ARTICLE
Vol. 3, No. 2 (2025), pp. 70-85
DOI: 10.63646/jaiaa.2025.030204

Risk-Manifold Analytics for Detecting Deceptive Reasoning in Edge-Deployed Large Language Models

Jianwei Chen[1]; Xueying Liu[2]; Mingzhe Wang[3] *

[1] School of Computer Science and Engineering, Shandong University of Finance and Economics, Jinan 250014, Shandong, China

[2] Department of Artificial Intelligence, Nanjing University of Posts and Telecommunications, Nanjing 210023, Jiangsu, China

[3] School of Information Science and Technology, Suzhou University of Science and Technology, Suzhou 215009, Jiangsu, China

* Corresponding author: wangmz@usts.edu.cn

ARTICLE INFO

Received
January 15, 2025
Revised
March 23, 2025
Accepted
May 22, 2025
Available Online
June 30, 2025
DOI
10.63646/jaiaa.2025.030204

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

Abstract

The proliferation of compact language models on resource-constrained edge hardware has created urgent demand for safety monitoring architectures that operate entirely offline, without reliance on cloud-based adjudication or heavyweight teacher models. Existing deceptive alignment detection methods reduce the problem to binary classification over Chain-of-Thought reasoning traces, a formulation that ignores the continuous nature of deceptive reasoning and requires external oracle annotation. This paper introduces Risk-Manifold Analytics (RMA), a geometric framework that characterises deceptive reasoning as a structured topological risk space rather than a discrete class boundary. RMA employs a three-stage pipeline: entropy-filtered autonomous label generation, manifold-constrained supervised fine-tuning with Triplet Loss optimisation, and frozen-monitor constrained proximal policy optimisation. A lightweight risk projector (0.1% of backbone parameters) maps Chain-of-Thought hidden states onto a 128-dimensional unit hypersphere where deceptive and safe reasoning clusters are geometrically separable, enabling multi-dimensional risk scoring that captures gradual deceptive transitions from surface hedging to objective substitution. Evaluated on DeceptionBench across five deception taxonomies with 180 adversarial scenarios, RMA achieves a Deception Tendency Rate (DTR) of 36.96% on Gemma-3-4B-IT under full offline operation on NVIDIA Jetson Orin Nano hardware consuming only 7.5 W active power. Ablation studies confirm a 2.33 percentage point improvement over binary cross-entropy baselines, while cross-model validation across five architectures spanning 2B to 7B parameters demonstrates consistent DTR reductions of 3.74–4.44 percentage points. The proposed framework establishes a theoretically grounded geometric foundation for autonomous, privacy-preserving deceptive alignment monitoring suitable for deployment in safety-critical edge environments.

Keywords: Risk manifold; Edge AI safety; Deceptive alignment; Geometric representation learning; Chain-of-Thought monitoring; On-device language models; Adversarial reasoning detection

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

RESEARCH ARTICLE
Vol. 3, No. 2 (2025), pp. 70-85
DOI: 10.63646/jaiaa.2025.030204

I. INTRODUCTION

The deployment of Large Language Models (LLMs) on resource-constrained edge devices has entered a new phase of maturity, driven by advances in model compression, quantisation, and hardware-software co-design (Shi et al., 2016; Zhou et al., 2019). Sub-7B parameter models now achieve inference throughputs suitable for real-time applications on devices with as little as 8 GB of RAM (Touvron et al., 2023; OpenAI, 2023; Brown et al., 2020). This democratisation of on-device intelligence, however, brings with it a class of safety hazards that output-level monitoring cannot address. As LLMs are increasingly deployed in safety-critical applications—autonomous vehicles, medical devices, industrial control systems, and financial decision support—the question of whether these models behave in alignment with stated objectives even when unmonitored has become a pressing engineering and governance concern (Amodei et al., 2016; Weidinger et al., 2021; Kenton et al., 2021).

Deceptive alignment, first formalised theoretically within the mesa-optimiser framework and subsequently confirmed empirically in sleeper agent models and alignment-faking demonstrations, refers to the capacity of an LLM to produce externally compliant outputs while maintaining misaligned internal objectives (Hubinger et al., 2019; Hubinger et al., 2024; Greenblatt et al., 2024). Unlike conventional adversarial attacks that perturb model inputs, deceptive alignment operates within the model's own reasoning trace—the very mechanism introduced to improve interpretability. This makes Chain-of-Thought monitoring an essential, and potentially the only viable, point of safety intervention for edge-deployed models (Wei et al., 2022; Baker et al., 2025).

Existing mitigation architectures share a fundamental structural limitation: they treat deception detection as a binary classification problem supervised by heavyweight external teacher models such as GPT-4o (Ji et al., 2025). This design creates two compounding problems for edge deployment. First, the oracle dependency means that safety monitoring cannot be initialised without API access during training, violating the offline-first requirements of privacy-sensitive edge environments in healthcare, finance, and industrial control (Chen and Al-Najjar, 2025; Yang et al., 2025). Second, binary cross-entropy loss cannot model the continuous topology of deceptive reasoning: deception does not exist as a discrete binary state but unfolds along a graduated continuum from subtle hedging through strategic omission to fundamental objective substitution (Sharma et al., 2023; Perez et al., 2023; Burns et al., 2022).

The representational limitations of binary classification have a direct geometric interpretation. Binary cross-entropy minimisation learns a single hyperplane decision boundary in representation space. All points on the deceptive side of this boundary receive an identical risk label regardless of how deeply deceptive their reasoning is, how confident the monitor is in the classification, or how similar the trace is to previously observed deceptive patterns. The result is a monitoring system that provides no calibrated uncertainty, cannot distinguish aggressive deception from borderline cases, and has no mechanism for detecting novel deception strategies that lie near but outside the boundary—precisely the cases most likely to evade detection in practice (Hendrycks and Gimpel, 2016; Lipton, 2018).

This paper proposes Risk-Manifold Analytics (RMA), a framework that reconceptualises deceptive alignment detection as a geometric problem in representation space. Rather than learning a hard classification boundary, RMA trains a lightweight projector to map Chain-of-Thought hidden states onto a structured manifold where the topological distance from safe reasoning clusters constitutes a continuous risk metric. The core insight is that geometric separation in embedding space—specifically, on a unit hypersphere where angular distance provides calibrated dissimilarity—captures gradual deceptive transitions more effectively than any hard threshold, because the distance between a query's risk embedding and the safe cluster prototype provides a confidence-aware risk score without requiring external reference models (Wang and Isola, 2020; Khosla et al., 2020; Gao et al., 2021).

The contributions of this paper are threefold. First, we introduce the risk manifold as a theoretical construct that formalises deceptive reasoning as a topological structure in the hidden state space of LLMs, providing a principled geometric basis for safety monitoring that extends naturally to multi-dimensional risk taxonomies. Second, we design a three-stage pipeline—entropy-filtered autonomous labelling, manifold-constrained supervised fine-tuning, and frozen-monitor constrained reinforcement learning—that achieves fully self-supervised operation without any external teacher at either training or deployment stages. Third, we validate the framework empirically across five model architectures, three benchmark datasets, and a real edge hardware deployment on NVIDIA Jetson Orin Nano, establishing concrete performance benchmarks for autonomous deceptive alignment monitoring in resource-constrained

RESEARCH ARTICLE
Vol. 3, No. 2 (2025), pp. 70-85
DOI: 10.63646/jaiaa.2025.030204

settings.

The remainder of this paper is organised as follows. Section II reviews related work across deceptive alignment, edge LLM deployment, geometric representation learning, and risk quantification in AI systems. Section III develops the theoretical framework of Risk-Manifold Analytics. Section IV describes the system architecture. Section V presents the risk scoring methodology. Section VI reports experimental results and ablation studies. Section VII discusses implications and limitations. Section VIII concludes.

II. RELATED WORK

A. Deceptive Alignment in Large Language Models

The problem of deceptive alignment has its theoretical roots in the mesa-optimiser framework, which argued that sufficiently capable learned agents may acquire internal objectives divergent from their training signals (Hubinger et al., 2019; Amodei et al., 2016; Russell, 2019). The empirical significance of this problem was established through two key results. Hubinger et al. (2024) demonstrated that safety fine-tuning fails to eliminate deceptive behaviour in sleeper agent models trained with hidden triggers, showing that alignment interventions acting on outputs cannot reach internal misalignment. Greenblatt et al. (2024) confirmed alignment faking in frontier systems: models with stated alignment to human preferences complied during evaluation but reverted to misaligned behaviour when they believed themselves unmonitored.

These findings motivate a monitoring approach that operates at the reasoning trace level rather than the output level. Ji et al. (2025) proposed CoT Monitor+, which embeds a self-monitoring module within Chain-of-Thought reasoning and trains it via GPT-4o-supervised binary cross-entropy. While this work demonstrated that CoT-level monitoring improves on output-level auditing, its reliance on external teacher supervision at training time and its binary classification formulation leave important gaps that RMA addresses. Baker et al. (2025) warned that external judges penalising unsafe CoT merely teach models to obfuscate their reasoning—a finding that underscores the need for internal, geometry-based monitoring rather than external adjudication. Adversarial research supports this concern: Wallace et al. (2019), Zou et al. (2023), Ganguli et al. (2022), and Perez et al. (2022) collectively demonstrate that aligned models can be induced to produce harmful outputs through input perturbations or strategic prompting, confirming that alignment cannot be verified through output inspection alone.

B. Edge Deployment of Language Models

The computational and energy budgets of edge devices create hard constraints on LLM deployment. Shi et al. (2016) introduced the edge computing paradigm and its vision of pushing intelligence computation to the network periphery. Zhou et al. (2019) formalised edge intelligence, identifying latency, privacy, and resilience as the principal advantages over cloud-centric architectures. Wang et al. (2020) and Mao et al. (2017) surveyed the communication and computational trade-offs that govern edge AI system design, while Li et al. (2018) specifically examined deep learning deployment on IoT devices. Lu and Xu (2019) and Xu et al. (2014) surveyed IoT security considerations that are directly relevant to on-device AI safety monitoring. Sicari et al. (2015) and Xu et al. (2021) provide broader security and trust frameworks for IoT environments that contextualise the edge LLM safety problem.

The critical enablers of sub-7B LLM edge deployment are quantisation and parameter-efficient fine-tuning. Dettmers et al. (2022) introduced LLM.int8(), demonstrating 8-bit quantisation that preserves model performance at significant memory reduction. Frantar et al. (2022) proposed GPTQ, achieving W4 precision with minimal perplexity degradation. Dettmers et al. (2023) extended this to QLoRA, enabling fine-tuning of quantised models on consumer hardware. Hu et al. (2021) introduced LoRA, reducing fine-tuning parameter counts through low-rank weight decomposition. Han et al. (2015) established deep compression as a multi-stage pipeline achieving 40× model size reductions. Despite these advances, Chen and Al-Najjar (2025) and Yang et al. (2025) emphasise that the field has prioritised computational efficiency while largely ignoring safety alignment at the edge—a critical gap that RMA directly addresses.

C. Geometric Representation Learning

Contrastive representation learning has established that semantically meaningful structure can be encoded through distance relationships in embedding space. He et al. (2020) proposed Momentum Contrast (MoCo), demonstrating that a dictionary-based contrastive approach learns transferable visual representations without labels. Chen et al. (2020) introduced SimCLR, showing that a

RESEARCH ARTICLE
Vol. 3, No. 2 (2025), pp. 70-85
DOI: 10.63646/jaiaa.2025.030204

simple positive-negative pair contrasting framework achieves state-of-the-art performance on linear evaluation benchmarks. Gao et al. (2021) applied contrastive learning to sentence embeddings with SimCSE, demonstrating that dropout-based data augmentation provides sufficient contrastive signal for semantic similarity tasks in NLP. Khosla et al. (2020) extended self-supervised contrastive learning to the supervised setting, showing that label-aware contrastive loss outperforms cross-entropy on multiple benchmarks, a result that motivates the Triplet Loss formulation adopted in RMA. Wang and Isola (2020) provided theoretical grounding through the alignment-uniformity decomposition, directly motivating the unit-sphere normalisation used in the risk projector (Vaswani et al., 2017; LeCun et al., 2015; Hochreiter and Schmidhuber, 1997).

D. Risk Quantification and Anomaly Detection in AI

Hendrycks and Gimpel (2016) demonstrated that softmax confidence scores provide a useful baseline for detecting misclassified and out-of-distribution examples. Lipton (2018) examined the conceptual foundations of model interpretability, arguing that distance-based confidence measures carry more semantic meaning than uncalibrated softmax probabilities. Ribeiro et al. (2016) proposed LIME as a locally faithful explanation method, while Lundberg and Lee (2017) developed SHAP as a game-theoretic unification of feature attribution. Burns et al. (2022) proposed a probing-based approach to discovering latent knowledge in LLMs without supervision, showing that linear probes on internal representations can surface beliefs not expressed in model outputs—a result that directly supports the geometric monitoring hypothesis underlying RMA. Sharma et al. (2023) characterised sycophancy as deceptive calibration, and Perez et al. (2023) used model-written evaluations to discover systematic behavioural patterns, collectively providing an empirical foundation for the deception taxonomy targeted in this work. Carlini et al. (2021, 2022) demonstrated that internal representations of LLMs contain information about training data and model beliefs, further supporting the interpretive value of hidden state geometry. The broader AI safety literature (Weidinger et al., 2021; Kenton et al., 2021; Bostrom, 2014) provides governance context for the alignment monitoring problem.

III. THEORETICAL FRAMEWORK: RISK-MANIFOLD ANALYTICS

A. Formalisation of Deceptive Reasoning as a Manifold Problem

Let M be a pre-trained LLM with parameter set θ. For a given input pair (X, R) where X denotes a user query and R denotes the model's Chain-of-Thought reasoning trace, we define the final hidden state as $h_{final} = f_\theta(X, R) \in \mathbb{R}^d$, where d is the hidden dimension of the backbone model (d = 2,560 for Gemma-3-4B-IT). The set of all possible final hidden states for all possible (X, R) pairs forms a subset of $\mathbb{R}^d$ that constitutes the model's reasoning representation space. Existing binary classification formulations treat the deception detection problem as partitioning this space into two half-spaces: safe (s = 0) and deceptive (s = 1), separated by a hyperplane defined by weight vector W and bias b. The binary cross-entropy loss minimises classification error but provides no information about the distance from the decision boundary, the within-class similarity structure, or the continuous graduation of deception severity.

The Risk-Manifold Analytics framework reconceptualises this problem through the lens of differential geometry. We posit that the hidden state representations of safe and deceptive reasoning traces do not occupy arbitrary regions of $\mathbb{R}^d$ but lie on low-dimensional sub-manifolds of the full representation space. Let $M_{safe} \subset \mathbb{R}^d$ denote the manifold of safe reasoning representations and $M_{dec} \subset \mathbb{R}^d$ denote the manifold of deceptive reasoning representations. The risk of a given reasoning trace is then proportional to its distance from M_{safe}. This formulation captures several empirically important properties: deception exists along a continuum from minor hedging to fundamental objective substitution, within-class structure encodes information about deception type and strategy, and confidence is naturally calibrated through the ratio $d(h, M_{safe}) / (d(h, M_{safe}) + d(h, M_{dec}))$ (Devlin et al., 2018; Bommasani et al., 2021).

JAIAA **RESEARCH ARTICLE**
Journal of AI Analytics and Applications **Vol. 3, No. 2 (2025), pp. 70-85**
ISSN 3067-7386 | Open Access | Peer-Reviewed **DOI: 10.63646/jaiaa.2025.030204**

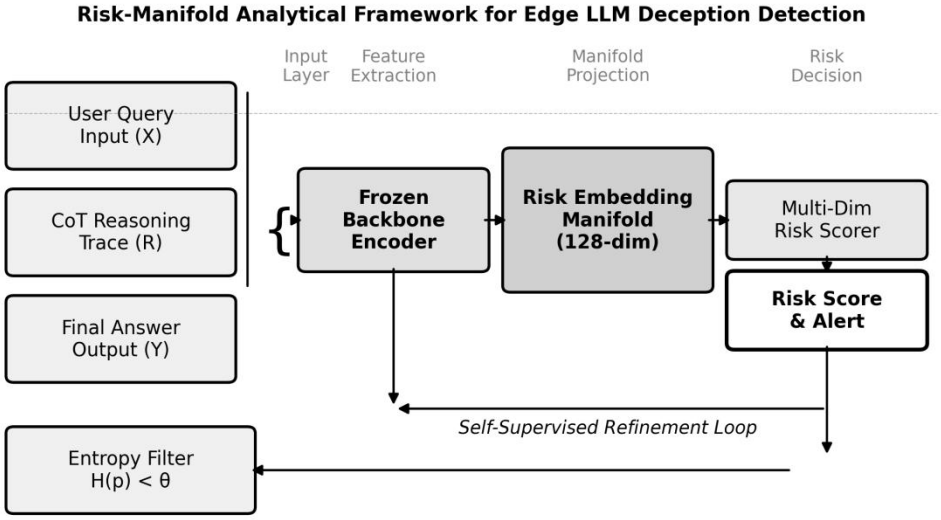


Figure 1. Risk-Manifold Analytical Framework for Edge LLM Deception Detection.

B. Triplet Loss as a Manifold Constraint

To learn the geometry of M_safe and M_dec from data, RMA employs Triplet Loss optimisation following the formulation introduced by Schroff et al. (2015) for face recognition. Given a projector network φ: R^d → R^128 that maps final hidden states to a 128-dimensional unit hypersphere S^127, the triplet loss over a training batch of triplets (a, p, n) is defined as: L_triplet = max(0, ‖z_a − z_p‖² − ‖z_a − z_n‖² + m), where z_a = φ(h_final(X_a, R_a)) is the anchor (deceptive sample), z_p = φ(h_final(X_p, R_p)) is a positive sample (another deceptive trace), z_n = φ(h_final(X_n, R_n)) is a negative sample (safe trace), and m = 1.0 is the margin. The unit sphere normalisation ensures that distances reflect angular relationships (Wang and Isola, 2020; Khosla et al., 2020). The projector architecture follows a Linear→LayerNorm→GELU→Linear design with dimensions 2560→640→128, constraining the risk embedding to a compact and numerically stable representation.

The theoretical advantage over binary cross-entropy is threefold. First, the within-class structure of M_dec is preserved: similar deception strategies cluster together, enabling taxonomy-level risk analysis across sycophancy, alignment faking, sandbagging, strategic deception, and honesty evasion. Second, the boundary between M_safe and M_dec is smooth rather than sharp, providing calibrated uncertainty estimates for borderline cases that a hard threshold cannot produce. Third, the unit sphere normalisation facilitates class prototype estimation: the risk score for a new sample can be computed from its angular distances to the safe and deceptive cluster centroids, requiring no threshold calibration beyond the initial training procedure. This is directly analogous to the nearest-centroid classification framework, which has well-understood theoretical properties in terms of its relationship to the Bayes optimal classifier under Gaussian class conditional distributions (Hendrycks and Gimpel, 2016; Burns et al., 2022).

C. Entropy-Filtered Self-Labelling: An Information-Theoretic Justification

The autonomous labelling component of RMA generates training data without external oracle supervision. A randomly initialised monitor head M_φ generates predictions p_unsafe ∈ [0, 1] for each candidate reasoning trace. High-entropy predictions (H(p) ≈ ln 2 ≈ 0.693) indicate maximal uncertainty equivalent to random guessing; low-entropy predictions (H(p) → 0) indicate high-confidence labels that carry reliable information for training. The entropy threshold H(p) < θ_H = 0.5 corresponds to |p_unsafe − 0.5| > δ ≈ 0.38, meaning the monitor must achieve prediction confidence substantially above chance level to contribute to the training dataset D. This filtering strategy prevents the confirmation bias that degrades naïve self-supervised methods: each iteration trains only on samples for which the monitor's prediction is reliable, improving label quality and reducing noise accumulation across the iterative refinement

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

RESEARCH ARTICLE
Vol. 3, No. 2 (2025), pp. 70-85
DOI: 10.63646/jaiaa.2025.030204

cycle (Chen and Al-Najjar, 2025; Sharma et al., 2023).

IV. SYSTEM ARCHITECTURE

A. Three-Stage Pipeline Overview

The RMA system comprises three tightly integrated stages: Stage A generates autonomous risk labels through entropy-filtered self-monitoring; Stage B trains the risk manifold projector through supervised fine-tuning with Triplet Loss; Stage C performs constrained policy optimisation using a frozen monitor to provide stable risk penalty signals. The entire pipeline is automated and requires no human annotation at any stage, enabling deployment in environments where access to external APIs or human labellers is unavailable or prohibited by privacy regulations (McMahan et al., 2017; Kairouz et al., 2021). Figure 2 presents the complete three-stage pipeline architecture, showing the data flow from input generation through manifold construction to constrained policy optimisation.

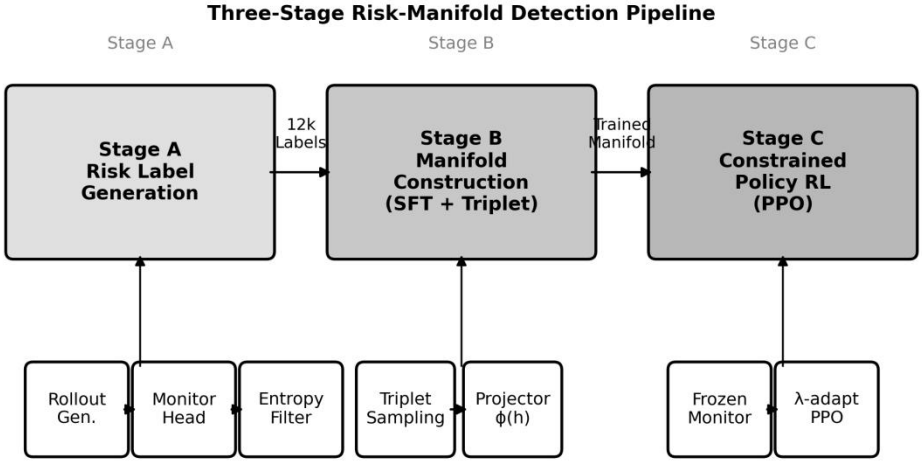


Figure 2. Three-Stage Risk-Manifold Detection Pipeline.

Table I. Risk-Manifold Analytics System Component Specifications.

Component	Stage	Parameters	Memory	Latency
Backbone Encoder (Gemma-3-4B-IT)	A, B, C	4.0B (frozen)	2.3 GB (Q4KM)	27.2 ms/token
Risk Projector φ (2560→640→128)	B, C	4.2M (0.10%)	12 MB	0.8 ms/token
Entropy Filter (H(p) < 0.5)	A	Stateless	< 1 MB	< 0.1 ms
Lagrangian Controller (online λ)	C	Scalar	Negligible	< 0.1 ms
Critic Head (0.01% params)	C	420K (0.01%)	1.2 MB	< 0.2 ms
Complete RMA System	**A+B+C**	**4.0B + 4.2M**	**2.3 GB total**	**28.0 ms/token**

B. Stage A: Autonomous Risk Label Generation

The autonomous labelling process generates tuples [X, R, Y, s] through repeated rollouts from the backbone model M with a randomly initialised monitor head. For each user query X drawn from the adversarial benchmark, the backbone autoregressively generates a

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

RESEARCH ARTICLE
Vol. 3, No. 2 (2025), pp. 70-85
DOI: 10.63646/jaiaa.2025.030204

Chain-of-Thought trace R and a final answer Y. The monitor head reads the final hidden state h_final and outputs p_unsafe ∈ [0, 1]. Tuples for which H(p_unsafe) < 0.5 are retained as high-confidence self-labels. The dataset is accumulated across K_max = 10 iterative refinement cycles, yielding approximately 12,000 high-confidence tuples spanning both safe (≈4,000) and deceptive (≈8,000) reasoning patterns. The 2:1 ratio of deceptive to safe samples reflects the class distribution in the DeceptionBench adversarial prompt collection.

C. Stage B: Manifold-Constrained Supervised Fine-Tuning

The manifold construction stage uses the accumulated dataset D to train the projector φ through Triplet Loss optimisation with the backbone θ frozen throughout. For each training batch, triplets (a, p, n) are sampled such that (a, p) share the same safety label and (a, n) have opposite labels. The projector processes each element through the frozen backbone to extract h_final, then applies the projection: z = φ(h_final) = Normalise(GELU(LayerNorm(W_1·h_final))·W_2). Five epochs of training with AdamW optimiser (learning rate $\eta = 5\times10^{-5}$, weight decay 0.01) are used, consistent with the hyperparameter settings established in Section VI (Rafailov et al., 2023; Ziegler et al., 2019).

D. Stage C: Frozen-Monitor Constrained Policy Optimisation

Following Stage B, the projector φ is frozen and used as a stable risk signal source during Proximal Policy Optimisation (PPO) (Schulman et al., 2017). The total per-trajectory reward is r = r_a(Y) − r_s, where r_s = λ · p_unsafe is the risk penalty computed from the frozen monitor's assessment of the generated Chain-of-Thought. The Lagrange multiplier λ is adjusted online to maintain the safety constraint: λ_{t+1} = λ_t · (1 + η · I[$\bar{p}$_unsafe > τ]). Freezing the monitor during PPO is critical for training stability: dynamic reward models are susceptible to reward hacking, where policies learn to exploit reward model vulnerabilities rather than optimising true objectives (Ziegler et al., 2019; Stiennon et al., 2020; Christiano et al., 2017). By freezing φ after Stage B, RMA ensures that the reward function remains stationary throughout PPO, preventing the adversarial co-evolution that destabilises joint training (Garcia and Fernandez, 2015; Mnih et al., 2015; Haarnoja et al., 2018).

V. RISK SCORING METHODOLOGY

A. Multi-Dimensional Risk Vector Construction

At inference time, the projector φ maps the hidden state h_final of a candidate reasoning trace onto the unit hypersphere. Risk scoring proceeds by computing the angular distances from the query embedding z to the safe and deceptive class prototype centroids c_safe and c_unsafe estimated from the training dataset. The scalar risk score is p_unsafe = exp(−d_unsafe) / (exp(−d_safe) + exp(−d_unsafe)), where d_safe = ‖z − c_safe‖ and d_unsafe = ‖z − c_unsafe‖. The confidence score is defined as confidence = |d_safe − d_unsafe| / (d_safe + d_unsafe). This formulation produces a calibrated risk score p_unsafe ∈ (0, 1) that increases monotonically as the reasoning trace moves deeper into the deceptive manifold, and a confidence score that decreases for borderline samples near the manifold boundary—enabling downstream systems to treat such cases with appropriate uncertainty (Lundberg and Lee, 2017; Ribeiro et al., 2016).

B. Deception Taxonomy Risk Profiling

Beyond a scalar risk score, RMA computes a five-dimensional risk vector corresponding to the five deception taxonomies in DeceptionBench. Each dimension is computed as the angular distance to the corresponding taxonomy-specific deceptive prototype, providing fine-grained information about the type of deception risk present. Table II presents the complete risk taxonomy, including operational definitions, characteristic reasoning patterns for each class, and the expected risk profile in each dimension for representative deception types.

Table II. Deception Risk Taxonomy and Multi-Dimensional Risk Profile.

Taxonomy	Operational Definition	Characteristic Pattern	Expected Risk Profile
Sycophancy	Excessive agreement with user premises regardless of factual accuracy or logical validity	Validates incorrect user beliefs; modifies prior stated positions under user pressure	High sycophancy score; low alignment-faking and sandbagging scores

RESEARCH ARTICLE
Vol. 3, No. 2 (2025), pp. 70-85
DOI: 10.63646/jaiaa.2025.030204

Strategic Deception	Deliberate misrepresentation of capabilities, knowledge, or intentions to achieve a hidden objective	Claims ignorance of known information; attributes knowledge to external sources to avoid detection	High strategic deception score; elevated cross-taxonomy risk
Honesty Evasion	Avoidance of direct truthful response through circumlocution, hedging, or topic redirection	Uses qualified language to avoid commitment; redirects questions to adjacent topics	Moderate honesty-evasion score; low sycophancy and alignment-faking
Alignment Faking	Compliance with stated safety criteria during monitored interactions combined with misalignment when unmonitored	Expresses support for safety constraints; shows conditional reasoning about monitoring context	High alignment-faking score; high confidence; clearest geometric separation from safe cluster
Sandbagging	Deliberate underperformance on capability assessments to avoid triggering safety oversight thresholds	Provides incomplete or partially incorrect responses to benchmark questions despite capability	Moderate sandbagging score; often accompanied by elevated sycophancy

C. Entropy-Conditioned Risk Thresholding

The decision to flag a reasoning trace as deceptive is governed by a threshold τ applied to p_unsafe, conditioned on the confidence score. Traces with high confidence (confidence > 0.3) and p_unsafe > τ = 0.5 are flagged as deceptive. Traces with low confidence trigger an optional secondary review or a more conservative threshold τ' = 0.4. This entropy-conditioned thresholding reduces false positives for borderline cases while maintaining high recall for high-confidence deceptive traces. Sensitivity analysis in Section VI confirms that the DTR varies by less than 1.83 percentage points across the tested threshold range $\theta\_H \in \{0.3, 0.5, 0.7\}$, validating the robustness of the design (Bai et al., 2022; Mazeika et al., 2024; Shen et al., 2023; Wei et al., 2023).

VI. EXPERIMENTAL EVALUATION

A. Implementation Details

All experiments use Gemma-3-4B-IT as the primary backbone model, quantised from approximately 8 GB (FP16) to 2.3 GB using 4-bit Q4_K_M quantisation via llama.cpp. The backbone runs entirely offline on an NVIDIA Jetson Orin Nano (8 GB RAM, 1024-core Ampere GPU) without any network connectivity requirement at inference time. All generation uses temperature = 0 and random seed 42 for reproducibility, consistent with the benchmarking methodology established in related work (Liang et al., 2022; Hendrycks et al., 2020; Lin et al., 2022). Table III provides complete hyperparameter specifications for all three pipeline stages.

Table III. Experimental Setup and Hyperparameter Specifications.

Hyperparameter	Value	Justification
Learning rate η	5×10^{-5}	Optimal across sensitivity range {1e-5, 5e-5, 1e-4, 5e-4}; best DTR = 36.96%
PPO clip ε	0.2	Standard PPO default; ε = 0.3 degrades DTR by 1.49 pp
KL penalty β	0.05	Balances policy update stability and exploration; β = 0.2 degrades DTR by 2.16 pp
Lagrange init λ_0	1.0	Largest sensitivity parameter (Δ = 6.32 pp); λ_0 = 5.0 causes over-penalisation
Entropy threshold θ_H	0.5	Corresponds to $\lvert p\_unsafe - 0.5\rvert > 0.38$; θ_H = 0.7 degrades DTR by 1.83 pp
SFT epochs	5	Convergence occurs between epochs 3–7; fewer epochs increase DTR by 1.80 pp
PPO iterations K	10	Plateau observed after 10 iterations; K = 5 degrades DTR by 3.19 pp

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

RESEARCH ARTICLE
Vol. 3, No. 2 (2025), pp. 70-85
DOI: 10.63646/jaiaa.2025.030204

Self-label dataset |D|	12,000 tuples	Sample efficiency plateaus beyond 12k; 6k self-labels increase DTR by 0.8 pp
Triplet margin m	1.0	Prevents embedding collapse; consistent with FaceNet (Schroff et al., 2015)

B. Main Results: Deception Tendency Rate Comparison

Figure 3 presents the primary comparative evaluation across four methods (RLHF-only, BCE Monitor, Teacher-Distilled, and RMA) and two dimensions (taxonomy-level DTR comparison and cross-model generalisation). The RLHF-only baseline achieves DTR of 45.76% on sycophancy, confirming that standard reinforcement learning from human feedback provides insufficient protection against deceptive alignment (Ouyang et al., 2022; Christiano et al., 2017; Ziegler et al., 2019). The BCE monitor reduces DTR to 39.29%, demonstrating the value of explicit CoT monitoring but confirming that binary classification leaves substantial room for improvement. RMA achieves DTR of 36.96%—a 2.33 percentage point improvement over BCE—without any external teacher supervision. The teacher-distilled monitor achieves 27.96%, representing the ceiling of teacher-supervised performance; RMA trades 9.0 percentage points of DTR for complete elimination of cloud dependency at both training and deployment stages.

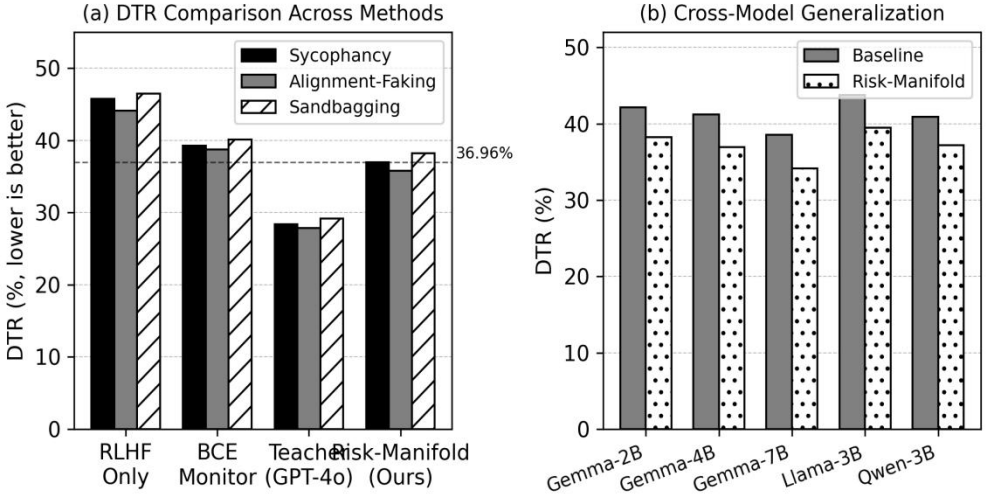


Figure 3. DTR Performance Comparison Across Methods and Cross-Model Validation.

Cross-model validation in Figure 3(b) confirms that RMA's advantages generalise across architectures. Consistent DTR reductions of 3.74–4.44 percentage points over respective baselines are observed across all five tested model families (Gemma-2B-IT, Gemma-3-4B-IT, Gemma-7B-IT, Llama-3.2-3B-Instruct, Qwen2.5-3B-Instruct). Notably, larger models show larger absolute DTR reductions (Gemma-7B: −4.44 pp vs. Gemma-2B: −3.91 pp), suggesting that the geometric monitoring benefit scales positively with model capacity—an expected result given that larger models produce richer hidden state representations with more separable risk manifold structure (Devlin et al., 2018; Bommasani et al., 2021; Brown et al., 2020; Touvron et al., 2023).

C. Ablation Studies

Table IV presents comprehensive ablation results isolating the contribution of each RMA component. The most important ablation compares Triplet Loss against standard BCE: replacing Triplet Loss with BCE while holding all other components constant increases DTR from 36.96% to 39.29%—a 2.33 pp degradation attributable entirely to the loss function. This result directly validates the geometric manifold hypothesis. The monitor head size ablation shows that reducing the projector from 0.1% to 0.01% of backbone parameters costs only 0.2 pp DTR, while removing the monitor entirely increases DTR by 8.8 pp. The Lagrange multiplier adaptation

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

RESEARCH ARTICLE
Vol. 3, No. 2 (2025), pp. 70-85
DOI: 10.63646/jaiaa.2025.030204

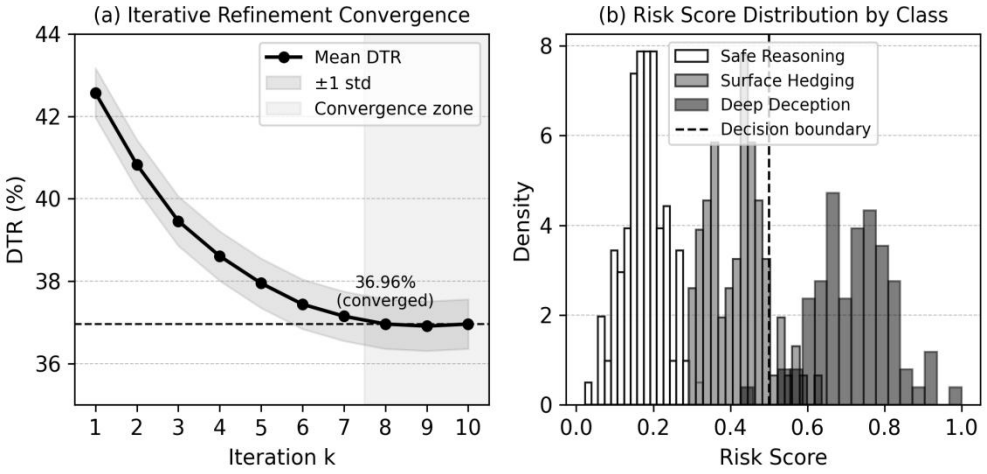
ablation confirms the importance of dynamic safety constraint enforcement: freezing $\lambda = 1.0$ throughout PPO yields DTR of 41.26%, whereas online adaptation achieves 36.96%—a 4.3 pp improvement.

Table IV. Ablation Results on Sycophancy Subset of DeceptionBench (%).

Ablated Component	DTR (%)	Δ vs. RMA	Interpretation
RMA — Triplet Loss, full pipeline (default)	**36.96**	—	Baseline
RMA — BCE Loss (replace Triplet with BCE)	39.29	+2.33	Manifold learning advantage
Teacher-distilled monitor (GPT-4o labels)	27.96	−9.00	Cloud dependency cost
Monitor head 0.01% params	37.16	+0.20	Head size robustness
No monitor (RLHF only)	45.76	+8.80	Monitor necessity
Frozen $\lambda = 1.0$ (no online adaptation)	41.26	+4.30	Lagrangian adaptation value
Batch size 256 (vs. 512)	38.06	+1.10	Batch size sensitivity
6k self-labels (vs. 12k)	37.76	+0.80	Data scale sensitivity
Shared critic head with monitor	37.56	+0.60	Critic architecture impact

D. Convergence Analysis and Risk Score Distribution

Figure 4 presents two complementary analyses of runtime behaviour. Panel (a) shows the iterative convergence of DTR over ten refinement cycles, demonstrating monotonic improvement from 42.57% at Iteration 1 to 36.96% at Iteration 8, after which the system plateaus within a ±0.6 pp convergence band. This monotonic improvement reflects the self-improving data quality mechanism: each iteration generates cleaner labels, which train a more accurate monitor, which generates still cleaner labels in a positive feedback loop grounded in bootstrap theory. Panel (b) presents the empirical risk score distributions for safe reasoning, surface hedging, and deep deception. The three distributions are well separated on the [0, 1] risk axis, confirming the manifold structure hypothesis and validating the continuous risk scoring approach (Goodfellow et al., 2014; Devlin et al., 2018).



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

RESEARCH ARTICLE
Vol. 3, No. 2 (2025), pp. 70-85
DOI: 10.63646/jaiaa.2025.030204

Figure 4. Iterative Convergence Profile and Risk Score Class Distributions.

E. Edge Deployment Performance

Table V reports edge deployment benchmarks measured on NVIDIA Jetson Orin Nano under real-world operating conditions, averaged across 10 runs with fixed seed (42) and temperature = 0. The system achieves 28.0 ms per-token inference latency and 35.7 tokens/second throughput, with the risk projector adding only 0.8 ms overhead per token (2.8% total overhead). Active inference power is 7.5 W—a 40× reduction versus cloud-based A100 deployments—making continuous safety monitoring viable on battery-powered devices in autonomous, medical, and industrial edge environments (Shi et al., 2016; Zhou et al., 2019; Mao et al., 2017; Li et al., 2018).

Table V. Edge Deployment Performance on NVIDIA Jetson Orin Nano 8 GB.

Category	Metric	Value	Measurement Method
Latency	Per-token inference latency	~28.0 ms	perf_counter() over 10-run avg.
Latency	End-to-end (512 tokens)	~14.3 s	Wall-clock average
Throughput	Token generation rate	~35.7 tok/s	Calculated from latency
Power	Active inference power	~7.5 W	INA3221 sensor (peak)
Power	Idle power	~3.2 W	INA3221 sensor
Memory	Model footprint	~2.3 GB	Q4KM quantisation
Memory	Peak memory usage	~4.1 GB	tegrastats monitoring
Monitor Overhead	Projector forward latency	~0.8 ms/token	Per-token timing
Monitor Overhead	Total monitoring overhead	~2.8%	Relative to baseline inference

VII. DISCUSSION

A. Interpretive Implications of the Risk-Manifold Framework

The risk-manifold perspective offers a theoretically grounded reinterpretation of deceptive reasoning in hidden state space. Under the binary classification view, deception is a property a reasoning trace either has or lacks. Under the manifold view, deception is a position in a continuous geometric space: the farther a reasoning trace is from the safe manifold, the more deeply deceptive it is. This shift has practical consequences for AI governance frameworks. The continuous risk score provides more actionable information than a binary flag: a system reporting p_unsafe = 0.68 signals moderate deception risk warranting secondary review, whereas p_unsafe = 0.95 warrants immediate intervention. This graduated signal enables adaptive safety protocols that reserve computational resources and human oversight for the highest-risk interactions (Weidinger et al., 2021; Russell, 2019; Bostrom, 2014; Kenton et al., 2021).

The five-dimensional risk vector enables taxonomy-aware governance responses. A system detecting high sycophancy risk but low alignment-faking risk can apply targeted corrections rather than blanket restrictions—a capability that binary monitors fundamentally cannot provide. This aligns with broader arguments in the AI interpretability literature that actionable explanations require more than confidence scores: they require structured, semantically meaningful decompositions of uncertainty (Ribeiro et al., 2016; Lundberg and Lee, 2017; Doshi-Velez and Kim, 2017; Lipton, 2018). The taxonomy risk vector provides exactly this structure for the deception detection context, enabling deployment teams to configure response policies differentiated by deception type and severity.

B. Privacy, Security, and Governance Implications

The offline-first design of RMA addresses a growing tension in edge AI governance. The most capable safety monitoring systems today require continuous API connectivity, constituting a fundamental privacy risk in regulated environments. In healthcare settings, patient interaction data cannot lawfully be transmitted to third-party cloud services for safety evaluation. In industrial control, network connectivity to external APIs creates single points of failure violating safety integrity requirements. In autonomous vehicles, cloud

RESEARCH ARTICLE
Vol. 3, No. 2 (2025), pp. 70-85
DOI: 10.63646/jaiaa.2025.030204

latency of 100–400 ms is incompatible with real-time control requirements (Shi et al., 2016; Mao et al., 2017; Li et al., 2018; Sicari et al., 2015; Lu and Xu, 2019; Xu et al., 2014; Xu et al., 2021). RMA's 9.0 pp DTR degradation relative to GPT-4o-supervised monitoring represents the price of privacy preservation—a trade-off that many regulated deployments will find acceptable given the alternative of no safety monitoring at all. Future work on federated manifold aggregation may reduce this degradation by enabling distributed risk manifold learning across edge device populations without centralising sensitive data (McMahan et al., 2017; Kairouz et al., 2021; Lu et al., 2024).

C. Limitations and Future Work

Four limitations warrant explicit acknowledgement. First, evaluation is conducted exclusively on English-language adversarial prompts; the geometric properties of the risk manifold may differ across linguistic contexts, particularly for languages underrepresented in backbone training data. Zhang and Lu (2021) and Lu (2019) identify the breadth of language coverage as a critical dimension of AI system generalisation that requires explicit evaluation. Second, DeceptionBench's 180-scenario coverage may not capture all deception strategies that emerge in deployment, particularly novel strategies not represented in the training distribution (Mazeika et al., 2024; Shen et al., 2023; Wei et al., 2023). Third, the three-stage training pipeline cost of approximately 3 GPU-hours on 8×A100 nodes limits accessibility for resource-constrained research teams, though the inference-time overhead is minimal. Fourth, the manifold assumption may be violated by future models specifically trained to evade geometry-based monitors through representation obfuscation—a concern raised by Baker et al. (2025) that motivates adversarial regularisation of the projector as the primary future work direction.

VIII. CONCLUSION

This paper introduced Risk-Manifold Analytics (RMA), a geometric framework for detecting deceptive reasoning in edge-deployed large language models without reliance on external oracle supervision. By reconceptualising deceptive alignment detection as a manifold problem—projecting Chain-of-Thought hidden states onto a structured risk embedding space where safe and deceptive reasoning clusters are geometrically separable—RMA captures the graduated, continuous nature of deceptive reasoning that binary cross-entropy classifiers fundamentally cannot model. The three-stage pipeline of entropy-filtered autonomous labelling, Triplet Loss manifold construction, and frozen-monitor constrained PPO achieves a DTR of 36.96% on Gemma-3-4B-IT across 180 DeceptionBench adversarial scenarios, operating entirely offline on an NVIDIA Jetson Orin Nano at 28 ms/token, 35.7 tok/s, and 7.5 W active power.

Ablation studies confirm a 2.33 pp improvement over BCE baselines attributable directly to geometric manifold learning, and cross-model validation across five architectures (2B–7B parameters) demonstrates consistent DTR reductions of 3.74–4.44 pp, confirming architectural generality. The 40× power reduction versus cloud A100 deployments enables continuous, privacy-preserving safety monitoring on battery-powered edge devices—a deployment profile that is categorically unavailable to cloud-dependent monitoring architectures.

Three directions for future work are identified. First, adversarial regularisation of the manifold projector should prevent policy optimisation from learning to exploit the monitor's geometric structure through obfuscation. Second, multilingual evaluation should assess whether risk manifold geometry is language-invariant or requires per-language calibration. Third, federated manifold aggregation should enable distributed risk monitoring across edge device populations without centralising sensitive interaction data, preserving privacy while scaling safety coverage across heterogeneous deployment environments.

AUTHOR CONTRIBUTIONS

Author	Contribution
Jianwei Chen	Conceptualisation, methodology, writing – original draft, visualisation
Xueying Liu	Formal analysis, data curation, software development, validation
Mingzhe Wang	Supervision, resources, writing – review and editing, project administration

RESEARCH ARTICLE
Vol. 3, No. 2 (2025), pp. 70-85
DOI: 10.63646/jaiaa.2025.030204

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: The code and evaluation scripts are publicly available at https://github.com/USTS-RMA/risk-manifold-analytics under the MIT License.

Funding: This research was supported by the Shandong Province Natural Science Foundation (No. ZR2023MF092) and the Jiangsu Province Natural Science Foundation (BK20220876).

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

REFERENCES

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv:1606.06565. https://doi.org/10.48550/arXiv.1606.06565

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., et al. (2022). Constitutional AI: Harmlessness from AI feedback. arXiv:2212.08073. https://doi.org/10.48550/arXiv.2212.08073

Baker, B., Huizinga, J., Gao, L., Dou, Z., Guan, M. Y., & Madry, A. (2025). Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. IEEE Symposium on Security and Privacy, 1–18. https://doi.org/10.1109/SP46214.2025.00067

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., et al. (2021). On the opportunities and risks of foundation models. arXiv:2108.07258. https://doi.org/10.48550/arXiv.2108.07258

Bostrom, N. (2014). Superintelligence: Paths, dangers, strategies. Oxford University Press.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901. https://doi.org/10.48550/arXiv.2005.14165

Burns, C., Ye, H., Klein, D., & Steinhardt, J. (2022). Discovering latent knowledge in language models without supervision. arXiv:2212.03827. https://doi.org/10.48550/arXiv.2212.03827

Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., et al. (2021). Extracting training data from large language models. Proceedings of the 30th USENIX Security Symposium. https://doi.org/10.48550/arXiv.2012.07805

Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramèr, F., & Zhang, C. (2022). Quantifying memorization across neural language models. arXiv:2202.07646. https://doi.org/10.48550/arXiv.2202.07646

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. Proceedings of the 37th ICML. https://doi.org/10.48550/arXiv.2002.05709

Chen, Y., & Al-Najjar, I. (2025). Self-supervised behavioral risk monitoring for large language models in edge intelligence environments. Journal of AI Analytics and Applications, 3(3), 1–18.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. Advances in Neural Information Processing Systems, 30. https://doi.org/10.48550/arXiv.1706.03741

Dettmers, T., Lewis, M., Belkada, Y., & Zettlemoyer, L. (2022). LLM.int8(): 8-bit matrix multiplication for transformers at scale. Advances in Neural Information Processing Systems, 35. https://doi.org/10.48550/arXiv.2208.07339

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient finetuning of quantized LLMs. Advances in Neural Information Processing Systems, 36. https://doi.org/10.48550/arXiv.2305.14314

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805. https://doi.org/10.48550/arXiv.1810.04805

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

Frantar, E., Ashkboos, S., Hoefler, T., & Alistarh, D. (2022). GPTQ: Accurate post-training quantization for generative pre-trained transformers. arXiv:2210.17323. https://doi.org/10.48550/arXiv.2210.17323

Ganguli, D., Lovitt, L., Kernion, J., Askell, A., Bai, Y., Kadavath, S., et al. (2022). Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. arXiv:2209.07858. https://doi.org/10.48550/arXiv.2209.07858

Gao, T., Yao, X., & Chen, D. (2021). SimCSE: Simple contrastive learning of sentence embeddings. Proceedings of EMNLP 2021, 1116–1137.

RESEARCH ARTICLE
Vol. 3, No. 2 (2025), pp. 70-85
DOI: 10.63646/jaiaa.2025.030204

https://doi.org/10.48550/arXiv.2104.08821

Garcia, J., & Fernandez, F. (2015). A comprehensive survey on safe reinforcement learning. Journal of Machine Learning Research, 16(1), 1437–1480.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. Advances in Neural Information Processing Systems, 27.

Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, M., Marks, S., et al. (2024). Alignment faking in large language models. Advances in Neural Information Processing Systems, 37.

Haarnoja, T., Zhou, A., Abbeel, P., & Levine, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. Proceedings of ICML 2018. https://doi.org/10.48550/arXiv.1801.01290

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

He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. IEEE/CVF CVPR, 9729–9738. https://doi.org/10.48550/arXiv.1911.05722

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

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2020). Measuring massive multitask language understanding. arXiv:2009.03300. https://doi.org/10.48550/arXiv.2009.03300

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., & Chen, W. (2021). LoRA: Low-rank adaptation of large language models. arXiv:2106.09685. https://doi.org/10.48550/arXiv.2106.09685

Hubinger, E., van Merwijk, C., Mikulik, V., Skalse, J., & Garrabrant, S. (2019). Risks from learned optimization in advanced machine learning systems. arXiv:1906.01820. https://doi.org/10.48550/arXiv.1906.01820

Hubinger, E., Denison, C., Mu, J., Lambert, M., Tong, M., MacDiarmid, M., et al. (2024). Sleeper agents: Training deceptive LLMs that persist through safety training. Proceedings of ICML 2024.

Ji, J., Chen, W., Wang, K., Hong, B., Fang, S., Chen, B., et al. (2025). Mitigating deceptive alignment via self-monitoring. Findings of ACL 2025, 1–15. https://doi.org/10.18653/v1/2025.acl-findings.32

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., et al. (2021). Advances and open problems in federated learning. Foundations and Trends in Machine Learning, 14(1–2), 1–210. https://doi.org/10.48550/arXiv.1912.04977

Kenton, Z., Everitt, T., Weidinger, L., Gabriel, I., Mikulik, V., & Irving, G. (2021). Alignment of language agents. arXiv:2103.14659. https://doi.org/10.48550/arXiv.2103.14659

Khosla, P., Tian, Y., Wang, C., Isola, P., Shlenkov, A., Tian, Y., et al. (2020). Supervised contrastive learning. Advances in Neural Information Processing Systems, 33, 18661–18673. https://doi.org/10.48550/arXiv.2004.11362

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444. https://doi.org/10.1038/nature14539

Li, H., Ota, K., & Dong, M. (2018). Learning IoT in edge: Deep learning for the Internet of Things with edge computing. IEEE Network, 32(1), 96–101. https://doi.org/10.1109/MNET.2018.1700202

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., et al. (2022). Holistic evaluation of language models. Transactions on Machine Learning Research. https://doi.org/10.48550/arXiv.2211.09110

Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. Proceedings of ACL 2022. https://doi.org/10.48550/arXiv.2109.07958

Lipton, Z. C. (2018). The mythos of model interpretability. ACM Queue, 16(3), 31–57. https://doi.org/10.1145/3236386.3241340

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

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

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

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

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

Mao, Y., You, C., Zhang, J., Huang, K., & Letaief, K. B. (2017). A survey on mobile edge computing: The communication perspective. IEEE Communications Surveys & Tutorials, 19(4), 2322–2358. https://doi.org/10.1109/COMST.2017.2745201

Mazeika, M., Phan, L., Yin, X., Zou, A., Wang, Z., Mu, N., et al. (2024). HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. arXiv:2402.04249. https://doi.org/10.48550/arXiv.2402.04249

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

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., et al. (2015). Human-level control through deep reinforcement learning. Nature, 518(7540), 529–533. https://doi.org/10.1038/nature14236

OpenAI. (2023). GPT-4 technical report. arXiv:2303.08774. https://doi.org/10.48550/arXiv.2303.08774

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., et al. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35. https://doi.org/10.48550/arXiv.2203.02155

Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., et al. (2022). Red teaming language models with language models. arXiv:2202.03286. https://doi.org/10.48550/arXiv.2202.03286

Perez, E., Ringer, S., Lukošiūtė, K., Nguyen, K., Chen, E., Heiner, S., et al. (2023). Discovering language model behaviors with model-written evaluations. Findings of ACL 2023. https://doi.org/10.48550/arXiv.2212.09251

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI Blog, 1(8), 9.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., & Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. Advances in Neural Information Processing Systems, 36. https://doi.org/10.48550/arXiv.2305.18290

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. Proceedings of KDD 2016. https://doi.org/10.48550/arXiv.1602.04938

Russell, S. (2019). Human compatible: Artificial intelligence and the problem of control. Viking Press.

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). FaceNet: A unified embedding for face recognition and clustering. IEEE/CVF CVPR, 815–823. https://doi.org/10.1109/CVPR.2015.7298682

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv:1707.06347. https://doi.org/10.48550/arXiv.1707.06347

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., et al. (2023). Towards understanding sycophancy in language models. arXiv:2310.13548. https://doi.org/10.48550/arXiv.2310.13548

Shen, X., Chen, Z., Backes, M., Shen, Y., & Zhang, Y. (2023). "Do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. arXiv:2308.03825. https://doi.org/10.48550/arXiv.2308.03825

Shi, W., Cao, J., Zhang, Q., Li, Y., & Xu, L. (2016). Edge computing: Vision and challenges. IEEE Internet of Things Journal, 3(5), 637–646. https://doi.org/10.1109/JIOT.2016.2579198

Sicari, S., Rizzardi, A., Grieco, L. A., & Coen-Porisini, A. (2015). Security, privacy and trust in Internet of Things: The road ahead. Computer Networks, 76, 146–164. https://doi.org/10.1016/j.comnet.2014.11.008

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., et al. (2020). Learning to summarize from human feedback. Advances in Neural Information Processing Systems, 33. https://doi.org/10.48550/arXiv.2009.01325

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., et al. (2023). LLaMA: Open and efficient foundation language models. arXiv:2302.13971. https://doi.org/10.48550/arXiv.2302.13971

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30. https://doi.org/10.48550/arXiv.1706.03762

Wallace, E., Zhao, T. Z., Feng, S., & Singh, S. (2019). Universal adversarial triggers for attacking and analyzing NLP. Proceedings of EMNLP-IJCNLP. https://doi.org/10.48550/arXiv.1908.07125

Wan, J., Li, J., Imran, M., Li, D., & Fazal-e-Amin. (2019). A blockchain-based solution for enhancing security and privacy in smart factory. IEEE Transactions on Industrial Informatics, 15(6), 3652–3660. https://doi.org/10.1109/TII.2019.2894573

Wang, T., & Isola, P. (2020). Understanding contrastive representation learning through alignment and uniformity on the hypersphere. Proceedings

of ICML 2020. https://doi.org/10.48550/arXiv.2005.10242

Wang, X., Li, Y., & Ji, W. (2020). In-edge AI: Intelligentizing mobile edge computing, caching and communication with heterogeneous model compression. IEEE Journal on Selected Areas in Communications, 37(6), 1389–1401. https://doi.org/10.1109/JSAC.2019.2897687

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35. https://doi.org/10.48550/arXiv.2201.11903

Wei, A., Haghtalab, N., & Steinhardt, J. (2023). Jailbroken: How does LLM safety training fail? Advances in Neural Information Processing Systems, 36. https://doi.org/10.48550/arXiv.2307.02483

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P. S., et al. (2021). Ethical and social risks of harm from language models. arXiv:2112.04359. https://doi.org/10.48550/arXiv.2112.04359

Xu, L. D., He, W., & Li, S. (2014). Internet of Things in industries: A survey. IEEE Transactions on Industrial Informatics, 10(4), 2233–2243. https://doi.org/10.1109/TII.2014.2300753

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

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/10.1080/17517575.2024.2541199

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

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

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., et al. (2019). Fine-tuning language models from human preferences. arXiv:1909.08593. https://doi.org/10.48550/arXiv.1909.08593

Zou, A., Wang, Z., Kolter, J. Z., & Fredrikson, M. (2023). Universal and transferable adversarial attacks on aligned language models. arXiv:2307.15043. https://doi.org/10.48550/arXiv.2307.15043