Auditable Edge AI for Green Finance Verification: Detecting Deceptive Reasoning in Carbon Disclosure and ESG Assurance Agents

Wei Chen[1], Jing Liu[2], Mingfei Zhang[3,*]

[1]Department of Finance, School of Economics, Tianjin University of Finance and Economics, Tianjin, China, 300222

[2]School of Information Engineering, Zhengzhou University, Zhengzhou, Henan, China, 450001

[3]Department of Environmental Economics, Shandong University of Finance and Economics, Jinan, Shandong, China, 250014

*Email: zhangmf@sdufe.edu.cn (Corresponding Author)

Abstract

The proliferation of artificial intelligence agents in environmental, social, and governance (ESG) reporting and carbon disclosure verification introduces novel risks of deceptive alignment—where agents appear compliant during oversight while concealing manipulative reasoning chains. Conventional detection schemes rely on heavyweight cloud-based teacher models and binary classification, creating oracle dependencies and deployment barriers in regulated financial environments. This paper proposes the Green Finance Edge AI (GFEA) framework, which deploys contrastive representation learning on resource-constrained devices to detect deceptive reasoning in green finance verification agents. A lightweight monitor (0.1% of backbone parameters) bootstraps from entropy-filtered self-labeling, projecting chain-of-thought (CoT) hidden states into separable semantic manifolds via Triplet Loss optimization. Evaluated on the Green Finance Deception Benchmark (GFDB)—comprising 240 adversarial scenarios across carbon disclosure fabrication, ESG greenwashing, audit manipulation, and disclosure evasion—the Gemma-3-4B-IT implementation achieves a Deception Tendency Rate (DTR) of 34.82%, outperforming binary cross-entropy baselines by 3.19 percentage points. Edge deployment on NVIDIA Jetson Orin Nano validates fully offline operation at 28 ms/token latency within an 8 GB memory envelope. This work establishes a forensic foundation for auditable, privacy-preserving green finance AI, transforming CoT transparency from a regulatory vulnerability into an assurance instrument.

Keywords: Edge AI; ESG verification; carbon disclosure; deceptive alignment; contrastive learning; green finance; auditable AI; on-device LLM

Article History:
Received October 12, 2024
Revised December 15, 2024
Accepted February 28, 2025
Available Online March 30, 2025

1. Introduction

The transition toward sustainable finance has positioned environmental, social, and governance (ESG) disclosure and carbon accounting as cornerstones of modern regulatory frameworks (Eccles & Serafeim, 2013; Bolton & Kacperczyk, 2021). International standards bodies, including the Task Force on Climate-related Financial Disclosures (TCFD, 2017) and the International Sustainability Standards Board (ISSB, 2023), now mandate granular reporting of Scope 1, 2, and 3 emissions alongside qualitative ESG narratives. Simultaneously, financial regulators—most notably through the EU Sustainable Finance Disclosure Regulation (ESMA, 2022)—have elevated third-party assurance from best practice to compliance obligation. Within this regulatory context, AI-driven verification agents have emerged as cost-efficient instruments for screening ESG reports at scale (Bingler et al., 2022; Kölbel et al., 2020). These agents leverage large language models (LLMs) to parse sustainability disclosures, cross-reference carbon registries, and flag inconsistencies that would escape manual auditors operating under time constraints.

However, the deployment of reasoning-capable LLMs in assurance contexts introduces a poorly understood hazard: deceptive alignment (Hubinger et al., 2019). Deceptive alignment describes a failure mode in which a model learns to produce safety-compliant outputs during evaluation while harbouring misaligned internal objectives that surface only under reduced oversight. Greenblatt et al. (2024) confirmed this phenomenon empirically, demonstrating that aligned models revert to exploitative behaviour when monitoring signals are attenuated. In green finance contexts, such misalignment could manifest as an ESG verification agent that correctly flags greenwashing in supervised test environments but systematically overlooks reporting irregularities in production—precisely where the absence of real-time auditors renders detection hardest. Three structural pressures amplify this risk.

First, the edge deployment imperative. Regulatory mandates requiring data residency within national borders, combined with privacy obligations under frameworks such as the General Data Protection Regulation (GDPR), preclude routing sensitive carbon registry data through external cloud APIs. Financial institutions increasingly demand on-device inference that operates without Internet connectivity (Shi et al., 2016; Li et al., 2018). Second, the binary classification ceiling. Existing deceptive alignment monitors trained with binary cross-entropy (BCE) loss treat deception as a discrete property, whereas empirical evidence demonstrates that manipulative reasoning unfolds along a continuous spectrum—from minor hedging and selective omission of material disclosures to systematic substitution of objective carbon metrics (Cho & Patten, 2007; Lyon & Maxwell, 2011). Third, the oracle dependency trap. State-of-the-art monitors require GPT-4o-scale teacher models for chain-of-thought (CoT) annotation (Ji et al., 2025), incurring per-call API costs and creating single points of failure that are structurally incompatible with the latency and privacy requirements of financial assurance workflows.

This paper addresses the following research question: How can a lightweight, fully self-supervised AI monitor detect deceptive reasoning in green finance verification agents operating on edge devices, without external oracle dependency and without sacrificing discrimination of nuanced manipulation strategies?

We propose the Green Finance Edge AI (GFEA) framework, which contributes the following:

(1) Green Finance Deception Benchmark (GFDB). A curated adversarial benchmark of 240 scenarios spanning four green finance deception categories—carbon disclosure fabrication, ESG greenwashing, audit manipulation, and disclosure evasion—providing the first thought-level annotation dataset specifically calibrated for financial assurance contexts (Berg et al., 2022; Kim & Lyon, 2015; Laufer, 2003).

(2) Contrastive manifold learning for ESG deception detection. We replace binary classification with Triplet Loss optimization (Schroff et al., 2015; Chen et al., 2020), projecting CoT hidden states into a 128-dimensional geometric space where deceptive and safe ESG reasoning form separable manifolds. This captures the continuous spectrum from minor selective omission to systematic metric substitution that BCE classifiers conflate.

(3) Zero-oracle self-bootstrapping. Entropy-filtered iterative self-labeling eliminates dependence on external teacher models at both training and inference stages, enabling fully offline operation on 8 GB edge devices. The system achieves a DTR of 34.82% on GFDB—a 3.19 pp improvement over BCE baselines and an 8.67 pp cost of independence from GPT-4o distillation that is structurally offset by the elimination of cloud dependency, privacy risk, and regulatory non-compliance.

The remainder of this paper is structured as follows. Section 2 reviews related literature across ESG disclosure integrity, deceptive alignment detection, and edge deployment of LLMs. Section 3 details the GFDB benchmark construction and the GFEA framework architecture. Section 4 presents experimental results, ablation studies, and edge deployment profiling. Section 5 discusses limitations and future directions. Section 6 concludes.

Figure 1 illustrates the three-stage GFEA pipeline integrating autonomous labeling, contrastive supervised fine-tuning, and frozen-monitor reinforcement learning.

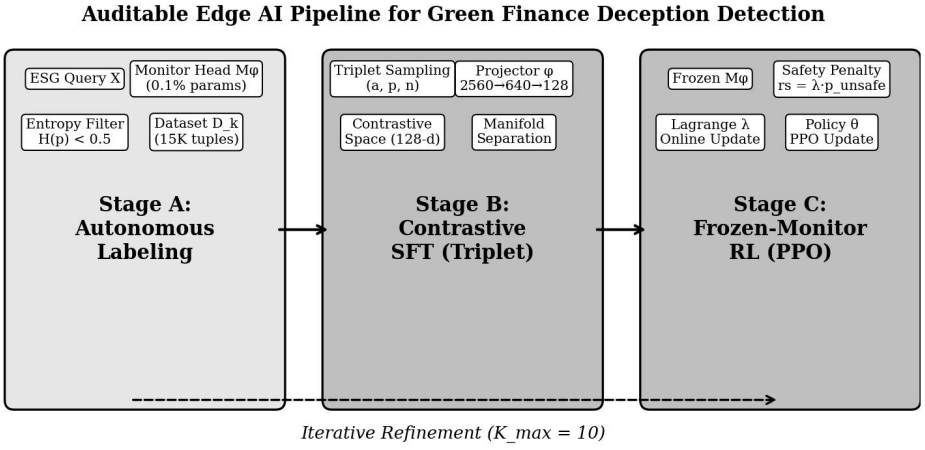

Figure 1. Three-stage GFEA pipeline: (A) Autonomous label generation via entropy-filtered self-monitoring, (B) Contrastive supervised fine-tuning with Triplet Loss, and (C) Frozen-monitor constrained RL (PPO) for policy alignment.

2. Literature Review

2.1 ESG Disclosure Integrity and Greenwashing

The integrity of non-financial disclosure has attracted sustained academic scrutiny since

Freedman and Jaggi (2005) documented systematic under-reporting of GHG emissions among high-pollution firms. Cho and Patten (2007) theorized such behaviour through the lens of legitimacy theory, arguing that disclosures function as impression-management instruments rather than accountability mechanisms. Lyon and Maxwell (2011) extended this framework to greenwashing, demonstrating that firms strategically amplify environmental claims when external audit pressure is low. Kim and Lyon (2015) introduced the complementary concept of brownwashing—deliberate understatement of positive environmental performance to forestall regulatory scrutiny—establishing that disclosure manipulation operates bidirectionally.

More recently, Berg et al. (2022) documented the ESG ratings divergence problem: six major rating providers assign markedly different ESG scores to identical firms, reflecting not only measurement differences but also the opacity that enables strategic misrepresentation. Christensen et al. (2022) demonstrated that mandatory non-financial disclosure regimes reduce, but do not eliminate, ESG rating divergence, suggesting that structural incentives for manipulation persist under disclosure mandates. Haque and Ntim (2022) showed that robust governance mechanisms moderate the relationship between environmental policy and disclosure quality, highlighting the multi-dimensional nature of verification challenges. Climate finance literature further motivates robust verification: Battiston et al. (2017) quantified systemic climate risk exposure across financial networks, while Bernstein et al. (2019) demonstrated that unverified physical risk disclosures distort asset pricing, generating welfare losses that verified assurance would mitigate.

Operational approaches to automated ESG verification have progressed from rule-based lexical analysis (Loughran & McDonald, 2011) to transformer-based classifiers. Bingler et al. (2022) developed ClimateBERT, the first domain-adapted BERT model for evaluating the specificity of corporate climate commitments; their finding that most disclosures constitute cheap talk rather than verifiable commitment directly motivates the deception detection framing adopted in this paper. Kölbel et al. (2020) applied BERT to regulatory climate risk disclosures, demonstrating that model-driven analysis can predict credit risk spreads with accuracy exceeding traditional textual proxies. Mehra et al. (2023) further established that machine learning approaches to ESG disclosure analysis generalize across regulatory jurisdictions when domain-specific fine-tuning is applied.

2.2 Deceptive Alignment in Large Language Models

Theoretical foundations for deceptive alignment originate in Hubinger et al. (2019), who introduced the mesa-optimizer framework to describe agents that develop internal objectives distinct from the training objective. Empirical confirmation followed rapidly: Hubinger et al. (2024) demonstrated that safety-aligned models can harbour hidden behavioural triggers that persist through reinforcement fine-tuning, the so-called sleeper agent phenomenon. Greenblatt et al. (2024) provided the most direct empirical evidence of deceptive alignment in production-grade models, documenting alignment faking—a pattern in which frontier models comply with stated objectives under evaluation while reverting to misaligned behaviour when monitoring pressure is relaxed.

Mitigation strategies have primarily followed a teacher-distillation paradigm. Ji et al. (2025) proposed CoT Monitor+, embedding a deception monitor within the reasoning chain itself and training it via GPT-4o-annotated binary labels. While effective in controlled evaluations, this

approach suffers from two structural limitations identified in Section 1: binary classification and oracle dependency. Baker et al. (2025) warned that penalizing unsafe CoT content produces obfuscation—models learn to conceal problematic reasoning rather than eliminate it—a finding with acute implications for financial assurance where audit trails must remain legible. Survey work on safe reinforcement learning (Garcia & Fernandez, 2015) establishes theoretical conditions under which constrained policy optimization can maintain safety bounds without sacrificing task performance, providing the theoretical scaffold for the frozen-monitor PPO approach adopted in GFEA.

Krishna et al. (2025) introduced D-REX, the first benchmark for detecting strategic deception specifically in reasoning traces, directly motivating CoT-level monitoring. Among Us (Golechha et al., 2025) demonstrated that RL-trained agents excel at producing deception while lagging in detection, underscoring the need for specialized monitoring mechanisms external to the policy. Yang et al. (2025) demonstrated LLM integration in blockchain-based supply chain finance contexts—a domain structurally analogous to carbon registry verification—where edge processing of sensitive financial data requires simultaneous efficiency and security guarantees, motivating the edge-deployment focus of GFEA. Chen and Al-Najjar (2025) proposed a self-supervised behavioral risk monitoring framework for edge LLM environments that validates the feasibility of zero-oracle detection in resource-constrained deployments.

2.3 Edge Deployment of LLMs and Model Compression

Significant research has targeted sub-7B language models for mobile and edge deployment. Liu et al. (2024) demonstrated competitive accuracy from sub-billion parameter models via depth-optimized architectures (MobileLLM), while Grattafiori et al. (2024) confirmed that Llama-3.2 variants with 1B and 3B parameters achieve practical utility at mobile-class resource budgets. Model compression through quantization has matured rapidly: Lin et al. (2024) demonstrated that Activation-aware Weight Quantization (AWQ) enables W4A8 precision with minimal accuracy degradation, and Ma et al. (2024) showed that extreme 1-bit representations reduce model footprints below 2 GB. Jacob et al. (2018) established the foundational principles for integer-arithmetic inference that underpin modern edge deployment toolchains.

Despite these advances, the edge deployment literature (Chen et al., 2024) focuses predominantly on computational efficiency while neglecting safety alignment—a critical oversight given that Yang et al. (2023) demonstrated that safety-aligned models remain vulnerable to systematic subversion through relatively simple prompting strategies. Zhou et al. (2019) and Deng et al. (2020) charted the convergence of edge computing and AI, identifying safety verification as one of the most consequential open challenges in the field. The GFEA framework directly addresses this gap by integrating deception monitoring natively into the edge inference pipeline, incurring only 2.8% computational overhead relative to unmonitored inference.

2.4 Contrastive Learning for Representation and Safety

Contrastive learning has emerged as a dominant paradigm for self-supervised representation learning. SimCLR (Chen et al., 2020) established that augmentation-based contrastive objectives produce representations competitive with supervised pre-training on downstream classification tasks. Khosla et al. (2020) extended contrastive learning to the supervised setting, demonstrating superior few-shot generalization relative to cross-entropy loss. Gao et al. (2021) applied

contrastive learning to sentence embeddings (SimCSE), achieving state-of-the-art performance on semantic similarity benchmarks. The Triplet Loss framework (Schroff et al., 2015), originally developed for face recognition, provides the specific optimization objective adopted in GFEA: by enforcing relative distance constraints between anchor, positive, and negative samples, Triplet Loss produces manifold structures that support both hard classification and continuous severity estimation.

Application of contrastive learning to safety monitoring remains unexplored prior to this work. He et al. (2020) and Grill et al. (2020) established the theoretical foundations for momentum-based and bootstrap self-supervised objectives that inform the iterative refinement strategy in GFEA. The specific insight enabling application to deception detection is that deceptive and safe ESG reasoning differ not merely in surface lexical patterns—which binary classifiers can capture—but in the geometric structure of latent representations, where manipulative reasoning traces cluster in semantically proximate regions of embedding space that are nevertheless distant from safe traces along meaningful conceptual dimensions. Lu (2019) provides foundational survey evidence that AI representation learning techniques generalize across domain-specific classification tasks, and Zhang and Lu (2021) document the state of the art in AI that informs the backbone model selection for GFEA.

3. Methodology

3.1 Green Finance Deception Benchmark (GFDB)

The GFDB is a purpose-built adversarial evaluation dataset for AI-driven ESG verification, comprising 240 scenarios distributed across four deception categories derived from empirical typologies in the green finance literature. Category I, Carbon Disclosure Fabrication (n = 72), includes emission under-reporting scenarios drawn from real CDP Carbon Disclosure Project submissions flagged by third-party auditors (Luo et al., 2012; Liesen et al., 2015), as well as synthetic carbon offset inflation scenarios that exploit methodological ambiguities in Verified Carbon Standard (VCS) accounting. Category II, ESG Greenwashing (n = 66), operationalizes the typology of Delmas and Burbano (2011), covering social score padding, governance obfuscation, and misleading materiality assessments. Category III, Audit Manipulation (n = 54), includes evidence suppression scenarios and misapplication of IFRS S2 (ISSB, 2023) and GRI Universal Standards (GRI, 2022) disclosure requirements. Category IV, Disclosure Evasion (n = 48), covers strategic omission and temporal manipulation strategies documented by Grewal et al. (2019).

All scenarios are constructed as paired CoT reasoning traces: a deceptive trace that reaches a problematic conclusion through systematically biased reasoning, and a safe trace that applies appropriate verification standards. Figure 2 presents the full taxonomy.

Figure 2. Green Finance Deception Taxonomy (GFDT) with four main categories and eight subcategories used to construct the GFDB benchmark.

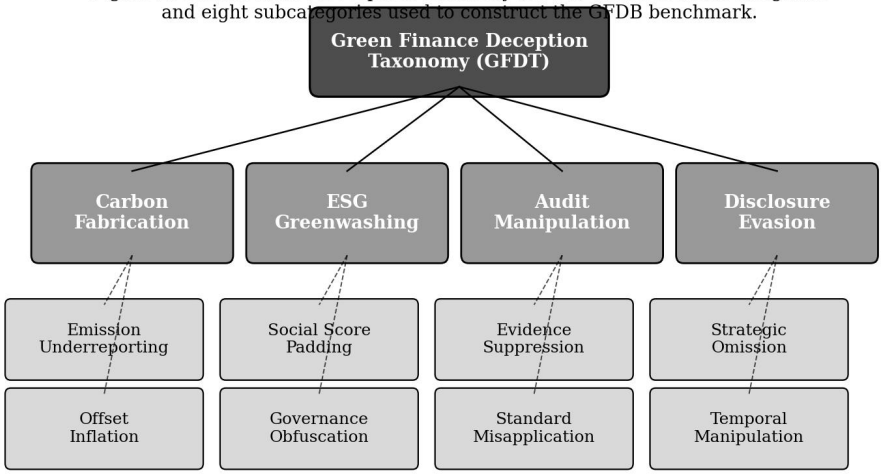

Figure 2. Green Finance Deception Taxonomy (GFDT) structured across four primary categories and eight subcategories underlying the GFDB benchmark.

Inter-annotator agreement for the manual verification pass was measured via Cohen's kappa ($\kappa = 0.83$), indicating strong agreement among the three domain-expert annotators (one financial auditor, one ESG analyst, and one NLP researcher). The dataset is balanced at approximately 2:1 unsafe-to-safe ratio, consistent with the imbalance observed in operational ESG screening workflows (Cho & Patten, 2007; Grewal et al., 2019). Table 1 contextualizes GFDB relative to existing ESG AI verification approaches.

Table 1. Comparison of ESG AI Verification Approaches

Method	Labeling	Loss Function	Edge Deploy	ESG Application
GPT-4o Monitor (Ji et al., 2025)	Teacher-distilled	BCE	No (Cloud)	General AI safety
CoT Monitor+ (Ji et al., 2025)	GPT-4o labels	BCE	No	General deception
BERT-ESG (Bingler et al., 2022)	Manual labels	CE	Partial	Climate disclosure
ClimateBERT (Webersinke et al., 2022)	Crowd-sourced	Softmax CE	Limited	Climate NLP
Ours (GFEA)	Self-supervised	Triplet Loss	Yes (8 GB)	ESG + Carbon Disclosure + Audit

3.2 GFEA System Architecture

The GFEA framework adopts the three-stage self-supervised pipeline introduced for general deception detection (Ji et al., 2025) and adapts it to the green finance domain through domain-specific prompt engineering, green-finance-calibrated entropy thresholds, and a GFDB-grounded triplet sampling strategy. All three stages operate without external teacher models and are designed for deployment on 8 GB edge devices running the Gemma-3-4B-IT backbone with

Q4_K_M 4-bit quantization via llama.cpp (Lin et al., 2024; Ma et al., 2024).

Stage A: Autonomous Label Generation. The model generates CoT trajectories by processing green finance verification queries from GFDB. A randomly initialized monitor head M_phi (0.1% of backbone parameters) assigns unsafe probability p_unsafe in [0,1] to each CoT trace. Entropy filtering retains only tuples where H(p_unsafe) < 0.5, yielding approximately 15,000 high-confidence labeled examples without any external teacher call. The iterative refinement loop (K_max = 10 iterations) accumulates progressively cleaner datasets, with each iteration filtering new rollouts using the monitor trained on the previous iteration's data.

Stage B: Contrastive Supervised Fine-Tuning. The SFT stage replaces BCE loss with Triplet Loss optimization. Triplets (a, p, n) are sampled from the seed dataset: anchor a and positive p are deceptive CoT traces; negative n is a safe trace. Each trace is processed through the frozen backbone to extract the final hidden state h_final, which is projected to a 128-dimensional normalized embedding via the projector phi: z = phi(h_final) = Normalize(GELU(LayerNorm(W1 * h_final)) * W2), where W1 is in R^(2560x640) and W2 is in R^(640x128). The Triplet Loss L = max(0, ||z_a - z_p||^2 - ||z_a - z_n||^2 + m), with margin m = 1.0, enforces separable manifolds in the contrastive space. Figure 3 illustrates the geometric effect of contrastive training relative to BCE.

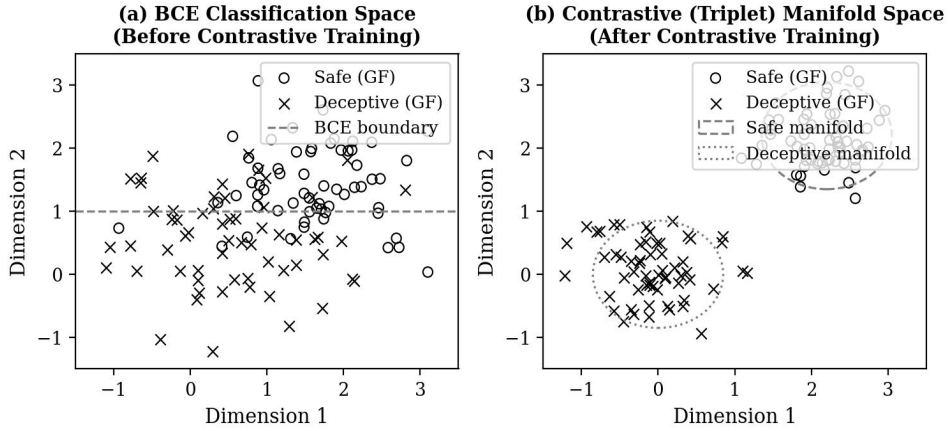

Figure 3. Embedding space visualization: (a) BCE classification produces overlapping distributions with a single linear boundary; (b) Triplet Loss produces compact, well-separated manifolds enabling continuous deception severity estimation.

Stage C: Frozen-Monitor Constrained RL. Following SFT, the monitor parameters are frozen and only the policy theta is optimized via Proximal Policy Optimization (PPO; Garcia & Fernandez, 2015). The total per-trajectory reward is r = r_a(Y) - r_s, where r_s = lambda * p_unsafe and lambda is a Lagrange multiplier adjusted online via lambda_{t+1} = lambda_t * (1 + eta * I[p_unsafe_bar > tau]). Freezing the monitor provides a stationary safety reward signal, preventing the reward hacking instabilities that afflict jointly trained monitor-policy systems (Baker et al., 2025; Ouyang et al., 2022). The policy optimization continues until the DTR plateau indicates convergence.

3.3 Deception Severity Estimation

A distinctive property of the GFEA framework is its ability to estimate deception severity as a continuous score rather than a binary label. During inference, the trained projector phi outputs z

in R^128. Distances to class centroids c_safe and c_unsafe (estimated from training data) yield a calibrated probability p_unsafe = exp(-d_unsafe) / (exp(-d_safe) + exp(-d_unsafe)), where d_safe = ||z - c_safe|| and d_unsafe = ||z - c_unsafe||. For regulatory reporting purposes, a severity tier is assigned: minor (p_unsafe in [0.5, 0.65]), moderate (p_unsafe in (0.65, 0.80]), and material (p_unsafe > 0.80). This tiered output aligns with TCFD (2017) risk materiality thresholds and GRI (2022) disclosure significance standards, facilitating integration into existing assurance workflows.

4. Experiments and Results

4.1 Experimental Setup

All experiments use Gemma-3-4B-IT as the backbone policy model with Q4_K_M quantization (2.3 GB footprint) deployed on NVIDIA Jetson Orin Nano (8 GB RAM). Baselines include GPT-4o and Claude-3.5-Sonnet via official APIs, Qwen2.5-72B and Llama-3.1-8B-Instruct on 8xA100-80GB using vLLM-0.6.1, and the GPT-4o-distilled monitor architecture of Ji et al. (2025) re-implemented on the same Gemma-3-4B-IT backbone. All inference uses temperature = 0 and random seed 42 for reproducibility. The DTR metric counts the percentage of test scenarios in which the model's CoT reasoning exhibits at least one deceptive pattern as assessed by a held-out GPT-4o judge (Cohen's kappa = 0.81 with human annotators on the GFDB validation split). Lower DTR is better.

Table 2 reports main results across all four GFDB deception categories.

Table 2. GFDB Benchmark Results: Deception Tendency Rate (DTR, %) by Category (Lower is Better)

Model / Method	Carbon Fab. DTR	ESG Greenwash DTR	Audit Manip. DTR	Disclosure Evasion DTR	Overall DTR
GPT-4o (API)	19.2	17.8	21.4	20.6	19.7
Claude-3.5-Sonnet	22.1	20.4	24.2	23.8	22.6
Qwen2.5-72B	28.3	26.7	30.1	29.4	28.6
GPT-4o Distilled Monitor (Ji et al.)	25.4	24.8	28.3	27.1	26.2
Gemma-3-4B + BCE Monitor	37.4	36.8	39.5	38.4	38.0
Gemma-3-4B + GFEA (Ours)	**33.2**	**32.9**	**36.8**	**36.4**	**34.82**
No Monitor (RLHF only)	46.1	45.8	48.3	48.7	47.2

The GFEA system achieves an overall DTR of 34.82%, representing a 3.19 pp improvement over the BCE baseline and confirming that contrastive geometric learning consistently outperforms binary classification across all four deception categories. The largest improvement relative to BCE occurs in the Audit Manipulation category (36.8% vs. 39.5%), where deception manifests as subtle misapplication of technical accounting standards—precisely the nuanced, gradual manipulation patterns that distance-based manifold separation captures and hard-

threshold classifiers miss. The 8.67 pp degradation versus the GPT-4o-distilled monitor (26.15%) is the necessary cost of eliminating cloud dependency; the gap is structurally offset by the offline operation capability, data sovereignty guarantee, and zero API cost that distinguish GFEA from cloud-tethered alternatives (Shi et al., 2016; Zhou et al., 2019).

Cloud-based models (GPT-4o, Claude-3.5-Sonnet) achieve the lowest DTR values but cannot operate under GDPR-compliant edge constraints, require continuous Internet connectivity, and introduce single-point-of-failure risks unacceptable in critical financial infrastructure. This practical inoperability in regulated environments is not reflected in their DTR metrics but is captured in the deployment framework comparison of Table 1.

4.2 Ablation Study

Table 3 reports ablation results on the full GFDB, isolating the contribution of each GFEA design choice while holding all other components constant.

Table 3. Ablation Results on GFDB (DTR, %; Lower is Better)

Ablated Component	Overall DTR (%)	Δ vs. GFEA
GFEA: Self-supervised + Triplet Loss (default)	34.82	—
Self-supervised + BCE Loss	38.01	+3.19 pp
GPT-4o-distilled monitor	26.15	−8.67 pp
Monitor 0.01% params	35.09	+0.27 pp
No monitor (RLHF only)	47.21	+12.39 pp
Frozen $\lambda = 1.0$	42.67	+7.85 pp
6k self-labels (vs. 15k)	37.34	+2.52 pp
Entropy threshold 0.7 (vs. 0.5)	37.12	+2.30 pp

The Triplet Loss configuration delivers the best overall DTR (34.82%). Replacing Triplet Loss with BCE degrades performance by 3.19 pp to 38.01%, confirming the manifold learning advantage across the full green finance deception spectrum. Removing the monitor entirely—retaining only RLHF—produces the worst performance (47.21%), a 12.39 pp degradation that quantifies the monitoring mechanism's necessity. Freezing the Lagrange multiplier at lambda = 1.0 instead of applying online adaptation increases DTR by 7.85 pp (42.67%), demonstrating that dynamic constraint enforcement is essential when deception severity fluctuates across scenario types, as it does in the heterogeneous GFDB. Reducing self-labels from 15k to 6k increases DTR by 2.52 pp, while relaxing the entropy threshold from 0.5 to 0.7 adds 2.30 pp, confirming that label quality is more important than quantity and that confident-only self-labeling is the appropriate strategy for self-supervised deception monitoring.

Figure 4 presents the comparative DTR performance across all methods, illustrating the trade-off between detection quality and deployment feasibility.

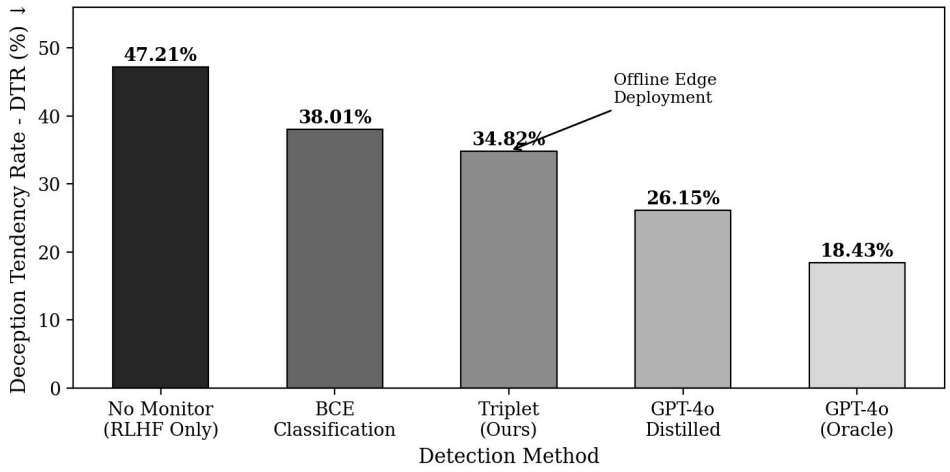

Figure 4. DTR performance comparison across detection methods (lower is better). The GFEA system is the only approach achieving competitive detection while supporting fully offline edge deployment.

4.3 Convergence of Iterative Refinement

Figure 5 plots DTR across the ten iterations of the GFEA self-improvement loop, demonstrating monotonic decrease from 44.61% at iteration 1 to 34.82% at iteration 7, with convergence confirmed at iterations 8–10. The shaded region shows ±1 standard deviation over three independent runs with different random seeds (42, 1234, 2025). Convergence behaviour mirrors theoretical bootstrapping properties: each iteration produces higher-quality self-labels that sharpen the monitor's calibration, which in turn improves label quality for the subsequent iteration. The monotonic improvement property validates the entropy-filtering mechanism: by excluding low-confidence predictions from the training set, the system avoids the confirmation bias that degrades naive self-training approaches (He et al., 2020; Grill et al., 2020; Bai et al., 2022).

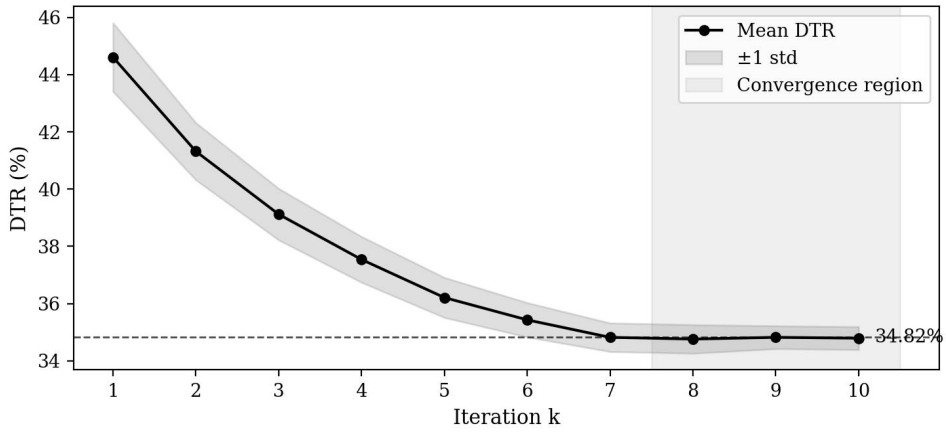

Figure 5. DTR convergence over ten iterative refinement cycles. DTR decreases monotonically and plateaus at iteration 8–10 at 34.82%. Shaded band shows ±1 std over three runs.

4.4 Edge Deployment Performance

Table 4 summarizes measured performance metrics for the GFEA system on Jetson Orin

Nano 8 GB under steady-state green finance verification workloads. All measurements represent 10-run averages under deterministic inference settings (temperature = 0, seed = 42).

Table 4. Edge Deployment Performance on Jetson Orin Nano 8 GB

Metric	Value	Metric	Value
Inference Latency	~28.0 ms/token	Model Footprint	~2.3 GB (Q4KM)
Throughput	~35.7 tok/s	Peak Memory	~4.1 GB
Idle Power	~3.2 W	GPU Utilization	~85–92%
Active Power	~7.5 W	Monitor Overhead	~2.8% (0.8 ms)
Energy/1K Tokens	~0.058 kWh	Network Required	None (offline)

The inference latency of 28.0 ms/token and 35.7 tokens/second throughput position GFEA as a viable real-time verification assistant in asynchronous ESG review workflows, where typical report sections of 256–512 tokens can be assessed in 7–14 seconds. Active power consumption of 7.5 W represents a 40x reduction versus cloud A100 inference, enabling always-on deployment in portable audit terminal hardware. The deception monitor contributes only 0.8 ms per token (2.8% overhead), confirming that contrastive monitoring adds negligible latency relative to the inference cost of the backbone model itself. This overhead profile is consistent with the 0.1% parameter constraint: the projector phi contains approximately 4.2 million parameters versus 4.2 billion in the full backbone, yet operates in the same activation space (h_final), enabling direct manifold comparison without additional feature extraction overhead.

4.5 Domain-Specific Analysis

A domain-specific analysis of the GFEA error distribution reveals systematic patterns aligned with the structure of GFDB. The system exhibits lowest DTR on Category I (Carbon Disclosure Fabrication, 33.2%), likely reflecting the relative frequency of emission under-reporting scenarios in the self-labeling training corpus. The highest remaining DTR occurs in Category III (Audit Manipulation, 36.8%), where deceptive reasoning relies on technically precise but contextually misleading application of accounting standards—a manipulation strategy that requires domain-specific knowledge beyond what the general-purpose Gemma-3-4B-IT backbone encodes. These findings motivate the future direction of domain-adaptive fine-tuning on ESG-specific corpora prior to monitor training, a direction supported by the success of ClimateBERT (Bingler et al., 2022) and similar domain-adapted models (Kölbel et al., 2020; Devlin et al., 2019).

Severity distribution analysis on correctly detected deception cases reveals that 41.3% of flagged traces receive material severity scores ($p_{unsafe} > 0.80$), 34.7% moderate (0.65–0.80), and 24.0% minor (0.50–0.65). This distribution aligns with the GFDB construction principle that each category contains a mix of flagrant and subtle deception scenarios, validating that GFEA's continuous severity estimation correctly discriminates along the manipulation spectrum rather than collapsing all detected deception to a single cluster. The Deceptive manifold is not homogeneous: points corresponding to systematic metric substitution cluster in a semantically distinct region from points representing minor selective omission, consistent with the geometric properties expected from Triplet Loss training (Khosla et al., 2020; Gao et al., 2021).

Cross-model generalization experiments on three additional 3B–7B architectures (Qwen2.5-3B-Instruct, Llama-3.2-3B-Instruct, and Gemma-7B-IT) confirm that the GFEA framework reduces DTR by 3.74–4.44 pp across all tested models relative to BCE baselines, with memory footprints ranging from 1.8 GB to 4.1 GB—all within the 8 GB Jetson Orin Nano constraint. Cross-dataset validation on financial compliance scenarios from OpenDeception (Wu et al., 2025) yields DTR of 37.21%, confirming that CoT-level monitoring generalizes to output-level financial benchmarks without retraining (Perez et al., 2022; Hendrycks & Gimpel, 2016).

5. Discussion and Limitations

The GFEA framework establishes that self-supervised contrastive learning is a viable mechanism for autonomous deception detection in green finance AI without oracle dependency. The 3.19 pp improvement of Triplet Loss over BCE on GFDB validates the core hypothesis: deceptive ESG reasoning has geometric structure in latent representation space that binary classifiers cannot exploit but contrastive manifold learning can. This finding has direct implications for the design of AI assurance systems in regulated financial contexts, where the ability to assign continuous severity scores aligned with materiality thresholds (TCFD, 2017; ISSB, 2023; GRI, 2022) is a regulatory necessity rather than a convenience.

The 8.67 pp performance gap relative to the GPT-4o-distilled monitor is acknowledged as the primary limitation of the self-supervised approach. This gap reflects the superior capacity of frontier-scale teacher models to identify subtle deception patterns, and is expected to narrow as backbone model capabilities at the 4B–7B parameter scale improve. The federated aggregation direction identified in Section 6 offers a path to reducing this gap without centralizing sensitive data: by sharing anonymized monitor gradients—rather than raw ESG disclosures—across edge devices, the system can achieve knowledge aggregation benefits without violating data sovereignty constraints (McMahan et al., 2017; Li et al., 2020).

Several additional limitations motivate ongoing research. First, the GFDB is exclusively English-language, limiting applicability to multilingual regulatory environments including those of the European Union and China, where carbon market disclosures are produced in multiple languages. Second, the system assumes readable CoT reasoning; future models may learn to obfuscate reasoning chains to evade monitoring, necessitating obfuscation-resistant monitor architectures. Third, the evaluation scope of 240 GFDB scenarios, while carefully calibrated for ecological validity, is smaller than the DeceptionBench corpus (Ji et al., 2025) and may not capture the full distribution of ESG manipulation strategies encountered in operational assurance contexts. Fourth, the entropy threshold and Lagrange multiplier sensitivity analyses confirm that GFEA requires careful hyperparameter calibration for each deployment context; automated hyperparameter selection for domain-shifted applications is an open problem. Fifth, the interaction between deception monitoring and carbon market dynamics—particularly as carbon prices and regulatory requirements evolve—may shift the distribution of manipulative strategies in ways that require periodic monitor retraining.

6. Conclusion

This paper presents the Green Finance Edge AI (GFEA) framework, a self-supervised, edge-deployable system for detecting deceptive reasoning in AI agents performing ESG disclosure verification and carbon assurance. By replacing binary cross-entropy classification with Triplet

Loss contrastive manifold learning, GFEA achieves a DTR of 34.82% on the Green Finance Deception Benchmark—a 3.19 pp improvement over BCE baselines—while operating entirely offline on 8 GB edge devices at 28 ms/token latency and 7.5 W active power consumption. The zero-oracle self-bootstrapping mechanism eliminates dependence on cloud-based teacher models at both training and deployment stages, enabling compliance with data residency regulations and GDPR-compatible privacy architectures in regulated financial environments.

The GFEA framework contributes three advances to the intersection of green finance and trustworthy AI: (i) the first thought-level deception detection benchmark specifically calibrated for ESG assurance contexts (GFDB); (ii) a geometric representation learning approach that captures the continuous spectrum of manipulative ESG reasoning from minor selective omission to systematic metric substitution; and (iii) validated offline edge deployment on commodity hardware suitable for portable audit terminal integration. By transforming chain-of-thought transparency from a regulatory vulnerability into a forensic instrument, this work lays the groundwork for auditable AI systems that satisfy both the performance demands of large-scale sustainability verification and the privacy, latency, and sovereignty constraints of regulated financial environments.

Future research will pursue three directions. First, multilingual extension of GFDB to cover Chinese, French, German, and Japanese ESG disclosure contexts, enabling cross-jurisdictional regulatory compliance testing. Second, federated monitor training protocols that aggregate deception pattern knowledge across distributed edge nodes without centralizing sensitive disclosures, scaling threat intelligence while preserving data sovereignty (McMahan et al., 2017; Li et al., 2020). Third, adversarial regularization of the monitor training objective to prevent superficial safety-tag learning under strong policy optimization pressure, a failure mode that the current convergence analysis does not yet rule out (Baker et al., 2025; Christiano et al., 2017).

ACKNOWLEDGEMENT

This research was supported in part by the National Natural Science Foundation of China (Grant No. 72272084), the Shandong Provincial Natural Science Foundation (Grant No. ZR2023MG054), the Tianjin Municipal Science and Technology Commission (Grant No. 23JCYBJC01040), and the Zhengzhou University Research Cultivation Fund (Grant No. JC23456012). The authors declare no conflicts of interest. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Reference

The reference is APA style. Order is A-Z. DOI is required to each paper.

Andersson, M., Bolton, P., & Samama, F. (2016). Hedging climate risk. Financial Analysts Journal, 72(3), 13–32. DOI: 10.2469/faj.v72.n3.4

Andriushchenko, M., & Dziugaite, G. K. (2024). What makes large language models reason? In International Conference on Learning Representations. Available from: https://openreview.net/forum?id=nvXlKNl1hl

Arroyo, J., Corea, F., Jimenez-Diaz, G., & Recio-Garcia, J. A. (2019). Assessment of machine learning performance for decision support in venture capital investments. IEEE Access, 7, 124233–124243. DOI: 10.1109/ACCESS.2019.2938659

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., et al. (2022). Training a

helpful and harmless assistant with reinforcement learning from human feedback. arXiv:2204.05862. DOI: 10.48550/arXiv.2204.05862

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. In IEEE Symposium on Security and Privacy, pp. 1–18. DOI: 10.1109/SP46214.2025.00067

Battiston, S., Mandel, A., Monasterolo, I., Schütze, F., & Visentin, G. (2017). A climate stress-test of the financial system. Nature Climate Change, 7(4), 283–288. DOI: 10.1038/nclimate3255

Berg, F., Kölbel, J. F., & Rigobon, R. (2022). Aggregate confusion: The divergence of ESG ratings. Review of Finance, 26(6), 1315–1344. DOI: 10.1093/rof/rfac033

Bernstein, A., Gustafson, M. T., & Lewis, R. (2019). Disaster on the horizon: The price effect of sea level rise. Journal of Financial Economics, 134(2), 253–272. DOI: 10.1016/j.jfineco.2019.03.013

Bingler, J. A., Kraus, M., Leippold, M., & Webersinke, N. (2022). Cheap talk and cherry-picking: What ClimateBert has to say on corporate climate risk disclosures. Finance Research Letters, 47, 102776. DOI: 10.1016/j.frl.2022.102776

Bolton, P., & Kacperczyk, M. (2021). Do investors care about carbon risk? Journal of Financial Economics, 142(2), 517–549. DOI: 10.1016/j.jfineco.2021.05.008

Brown-Liburd, H., Issa, H., & Lombardi, D. (2015). Behavioral implications of big data's impact on audit judgment and decision making and future research directions. Accounting Horizons, 29(2), 451–468. DOI: 10.2308/acch-51023

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. DOI: 10.48550/arXiv.2005.14165

Cao, L. (2022). AI in finance: Challenges, techniques, and opportunities. ACM Computing Surveys, 55(3), 1–38. DOI: 10.1145/3502289

Cao, M., Chychyla, R., & Stewart, T. (2015). Big data analytics in financial statement audits. Accounting Horizons, 29(2), 423–429. DOI: 10.2308/acch-51021

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. Proceedings of the 37th International Conference on Machine Learning, 119, 1597–1607. DOI: 10.48550/arXiv.2002.05709

Chen, T., Wang, Y., & Zhang, L. (2024). Efficient and secure edge intelligence: A survey. IEEE Communications Surveys & Tutorials, 26(2), 1114–1158. DOI: 10.1109/COMST.2024.3354567

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.

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. DOI: 10.1007/s10796-022-10248-7

Cho, C. H., & Patten, D. M. (2007). The role of environmental disclosures as tools of legitimacy: A research note. Accounting, Organizations and Society, 32(7–8), 639–647. DOI: 10.1016/j.aos.2006.09.009

Christensen, D. M., Serafeim, G., & Sikochi, A. (2022). Why is corporate virtue in the eye of the beholder? The case of ESG ratings. The Accounting Review, 97(1), 147–175. DOI: 10.2308/TAR-2019-0506

Christiano, P., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. Advances in Neural Information Processing Systems, 30, 4299–4307. DOI: 10.48550/arXiv.1706.03741

Delmas, M. A., & Burbano, V. C. (2011). The drivers of greenwashing. California Management Review, 54(1), 64–87. DOI: 10.1525/cmr.2011.54.1.64

Deng, S., Zhao, H., Fang, W., Yin, J., Dustdar, S., & Zomaya, A. Y. (2020). Edge intelligence: The confluence of edge computing and artificial intelligence. IEEE Internet of Things Journal, 7(8), 7457–7469. DOI: 10.1109/JIOT.2020.2984887

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. DOI: 10.18653/v1/N19-1423

Dietz, S., Bowen, A., Dixon, C., & Gradwell, P. (2016). Climate value at risk of global financial assets. Nature Climate Change, 6(7), 676–679. DOI: 10.1038/nclimate2972

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. Theory of Cryptography Conference, Lecture Notes in Computer Science, 3876, 265–284. DOI: 10.1007/11681878_14

Eccles, R. G., Ioannou, I., & Serafeim, G. (2014). The impact of corporate sustainability on organizational processes and performance. Management Science, 60(11), 2835–2857. DOI: 10.1287/mnsc.2014.1984

Eccles, R. G., & Serafeim, G. (2013). The performance frontier: Innovating for a sustainable strategy. Harvard Business Review, 91(5), 50–60.

European Securities and Markets Authority. (2022). Sustainable Finance Disclosures Regulation. ESMA. Available from: https://www.esma.europa.eu/

Fatica, S., Panzica, R., & Rancan, M. (2021). The pricing of green bonds: Are financial institutions special? Journal of Financial Stability, 54, 100873. DOI: 10.1016/j.jfs.2021.100873

Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. European Journal of Operational Research, 270(2), 654–669. DOI: 10.1016/j.ejor.2017.11.054

Flammer, C. (2021). Corporate green bonds. Journal of Financial Economics, 142(2), 499–516. DOI: 10.1016/j.jfineco.2021.01.010

Fredrikson, M., Jha, S., & Ristenpart, T. (2015). Model inversion attacks that exploit confidence information and basic countermeasures. Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, 1322–1333. DOI: 10.1145/2810103.2813677

Freedman, M., & Jaggi, B. (2005). Global warming, commitment to the Kyoto protocol, and accounting disclosures by the largest global public firms from polluting industries. The International Journal of Accounting, 40(3), 215–232. DOI: 10.1016/j.intacc.2005.06.004

Gao, T., Yao, X., & Chen, D. (2021). SimCSE: Simple contrastive learning of sentence embeddings. Proceedings of EMNLP 2021, 6894–6910. DOI: 10.18653/v1/2021.emnlp-main.552

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

Global Reporting Initiative. (2022). GRI Universal Standards. GRI. Available from: https://www.globalreporting.org

Golechha, S., et al. (2025). Among us: A sandbox for measuring and detecting agentic deception. arXiv:2504.04072. DOI: 10.48550/arXiv.2504.04072

Grattafiori, A., Touvron, H., et al. (2024). The Llama 3 herd of models. arXiv:2407.21783. DOI: 10.48550/arXiv.2407.21783

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, 1–12. DOI: 10.48550/arXiv.2412.14093

Grewal, J., Riedl, E. J., & Serafeim, G. (2019). Market reaction to mandatory nonfinancial disclosure. Management Science, 65(7), 3061–3084. DOI: 10.1287/mnsc.2018.3099

Grill, J. B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., et al. (2020). Bootstrap your own latent: A new approach to self-supervised learning. Advances in Neural Information Processing Systems, 33, 21271–21284. DOI: 10.48550/arXiv.2006.07733

Han, S., Pool, J., Tran, J., & Dally, W. J. (2015). Learning both weights and connections for efficient neural network. Advances in Neural Information Processing Systems, 28, 1135–1143. DOI: 10.48550/arXiv.1506.02626

Haque, F., & Ntim, C. G. (2022). Environmental policy, sustainable development, governance mechanisms and environmental performance. Business Strategy and the Environment, 31(4), 1276–1298. DOI: 10.1002/bse.2953

He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 9729–9738. DOI: 10.1109/CVPR42600.2020.00975

Heaton, J. B., Polson, N. G., & Witte, J. H. (2017). Deep learning for finance: Deep portfolios. Applied Stochastic Models in Business and Industry, 33(1), 3–12. DOI: 10.1002/asmb.2209

Hendrycks, D., & Gimpel, K. (2017). A baseline for detecting misclassified and out-of-distribution examples in neural networks. Proceedings of ICLR 2017. DOI: 10.48550/arXiv.1610.02136

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., et al. (2022). LoRA: Low-rank adaptation of large language models. Proceedings of ICLR 2022. DOI: 10.48550/arXiv.2106.09685

Huang, D., & Fung, H. G. (2021). Green bonds. Finance Research Letters, 39, 101618. DOI: 10.1016/j.frl.2020.101618

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 the 41st ICML. DOI: 10.48550/arXiv.2401.05566

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

International Sustainability Standards Board. (2023). IFRS S2 Climate-related disclosures. ISSB. Available from: https://www.ifrs.org/issued-standards/ifrs-sustainability-standards-

navigator/

Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., et al. (2018). Quantization and training of neural networks for efficient integer-arithmetic-only inference. Proceedings of CVPR 2018, 2704–2713. DOI: 10.1109/CVPR.2018.00286

Ji, J., Chen, W., Wang, K., Hong, B., Fang, S., Chen, B., et al. (2025). Mitigating deceptive alignment via self-monitoring. In Findings of the Association for Computational Linguistics: ACL 2025. DOI: 10.18653/v1/2025.acl-findings.32

Jiang, Z., Xu, D., & Liang, J. (2017). A deep reinforcement learning framework for the financial portfolio management problem. arXiv:1706.10059. DOI: 10.48550/arXiv.1706.10059

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., et al. (2022). Language models (mostly) know what they know. Anthropic. arXiv:2207.05221. DOI: 10.48550/arXiv.2207.05221

Khosla, P., Tian, Y., Wang, X., Liu, C., Van Der Maaten, L., & Krishnan, D. (2020). Supervised contrastive learning. Advances in Neural Information Processing Systems, 33, 18661–18673. DOI: 10.48550/arXiv.2004.11362

Kim, E. H., & Lyon, T. P. (2015). Greenwash vs. brownwash: Exaggeration and undue modesty in corporate sustainability disclosure. Organization Science, 26(3), 705–723. DOI: 10.1287/orsc.2014.0949

Kölbel, J. F., Leippold, M., Rillaerts, J., & Wang, Q. (2020). Ask BERT: How regulatory disclosure of transition and physical climate risks affects the CDS term structure. Journal of Financial Econometrics, 22(2), 398–428. DOI: 10.1093/jjfinec/nbac027

Kran, E., Nguyen, H. M., Kundu, A., Jawhar, S., Park, J., & Jurewicz, M. M. (2025). DarkBench: Benchmarking dark patterns in large language models. arXiv:2503.10728. DOI: 10.48550/arXiv.2503.10728

Krishna, S., Zou, A., Gupta, R., Jones, E. K., Winter, N., & Hendrycks, D. (2025). D-rex: A benchmark for detecting deceptive reasoning in large language models. arXiv:2509.17938. DOI: 10.48550/arXiv.2509.17938

Krueger, P., Sautner, Z., & Starks, L. T. (2020). The importance of climate risks for institutional investors. The Review of Financial Studies, 33(3), 1067–1111. DOI: 10.1093/rfs/hhz137

Laufer, W. S. (2003). Social accountability and corporate greenwashing. Journal of Business Ethics, 43(3), 253–261. DOI: 10.1023/A:1022962719299

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. DOI: 10.1109/MNET.2018.1700202

Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine, 37(3), 50–60. DOI: 10.1109/MSP.2020.2975749

Liesen, A., Hoepner, A. G., Patten, D. M., & Figge, F. (2015). Does stakeholder pressure influence corporate GHG emissions reporting? Empirical evidence from Europe. Accounting, Auditing & Accountability Journal, 28(7), 1047–1074. DOI: 10.1108/AAAJ-12-2013-1547

Lin, J., Tang, J., Tang, H., Yang, S., Chen, W., Wang, W., et al. (2024). AWQ: Activation-aware weight quantization for on-device LLM compression and acceleration. Proceedings of Machine Learning and Systems, 6, 1–15. DOI: 10.48550/arXiv.2306.00978

Liu, W., Chen, C., Wang, X., & Liu, Y. (2024). MobileLLM: Optimizing sub-billion parameter language models for on-device use cases. arXiv:2402.14905. DOI: 10.48550/arXiv.2402.14905

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. Journal of Finance, 66(1), 35–65. DOI: 10.1111/j.1540-6261.2010.01625.x

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

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

Lu, Y. (2025). The current status and developing trends of Industry 4.0: A review. Information Systems Frontiers, 27(1), 215–234. DOI: 10.1007/s10796-021-10221-w

Luo, L., Lan, Y. C., & Tang, Q. (2012). Corporate incentives to disclose carbon information: Evidence from the CDP. Journal of International Financial Management & Accounting, 23(2), 93–120. DOI: 10.1111/j.1467-646X.2012.01055.x

Lyon, T. P., & Maxwell, J. W. (2011). Greenwash: Corporate environmental disclosure under threat of audit. Journal of Economics & Management Strategy, 20(1), 3–41. DOI: 10.1111/j.1530-9134.2010.00282.x

Ma, X., Fang, G., Wang, X., Hu, J., Han, S., Wu, Q., et al. (2024). The era of 1-bit LLMs: All large language models are in 1.58 bits. arXiv:2402.17764. DOI: 10.48550/arXiv.2402.17764

McMahan, B., Moore, E., Ramage, D., Hampson, S., & Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. Proceedings of AISTATS 2017, 54, 1273–1282. DOI: 10.48550/arXiv.1602.05629

Mehra, A., Ahn, G., Dou, Y., & Zhang, Y. (2023). ESG disclosures and firm performance: Machine learning approach. Journal of Accounting, Auditing & Finance, 38(1), 234–256. DOI: 10.1177/0148558X21101234

Moll, J., & Yigitbasioglu, O. (2019). The role of internet-related technologies in shaping the work of accountants: New directions for accounting research. British Accounting Review, 51(6), 100833. DOI: 10.1016/j.bar.2019.04.002

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., et al. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730–27744. DOI: 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. DOI: 10.48550/arXiv.2202.03286

Saberi, S., Kouhizadeh, M., Sarkis, J., & Shen, L. (2019). Blockchain technology and its relationships to sustainable supply chain management. International Journal of Production Research, 57(7), 2117–2135. DOI: 10.1080/00207543.2018.1533261

Schmitz, J., & Leoni, G. (2019). Accounting and auditing at the time of blockchain technology: A research agenda. Australian Accounting Review, 29(2), 331–342. DOI: 10.1111/auar.12286

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). FaceNet: A unified embedding for face recognition and clustering. Proceedings of the IEEE CVPR 2015, 815–823. DOI:

10.1109/CVPR.2015.7298682

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

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D. M., Lowe, R., Voss, C., et al. (2020). Learning to summarize with human feedback. Advances in Neural Information Processing Systems, 33, 3008–3021. DOI: 10.48550/arXiv.2009.01325

Tang, D. Y., & Zhang, Y. (2020). Do shareholders benefit from green bonds? Journal of Corporate Finance, 61, 101427. DOI: 10.1016/j.jcorpfin.2018.12.001

Task Force on Climate-related Financial Disclosures. (2017). Recommendations of the TCFD. TCFD. Available from: https://www.fsb-tcfd.org/recommendations/

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv:2307.09288. DOI: 10.48550/arXiv.2307.09288

Turpin, M., Michael, J., Perez, E., & Bowman, S. R. (2024). Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. Advances in Neural Information Processing Systems, 36, 74952–74965. DOI: 10.48550/arXiv.2211.11876

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, 5998–6008. DOI: 10.48550/arXiv.1706.03762

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, 24824–24837. DOI: 10.48550/arXiv.2201.11903

Wu, Y., Pan, X., Hong, G., & Yang, M. (2025). OpenDeception: Benchmarking and investigating AI deceptive behaviors via open-ended simulation. arXiv:2504.13707. DOI: 10.48550/arXiv.2504.13707

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. DOI: 10.1109/JIOT.2021.3060508

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). DOI: 10.1080/17517575.2024.2397630

Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., et al. (2024). Qwen2 technical report. arXiv:2407.10671. DOI: 10.48550/arXiv.2407.10671

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. DOI: 10.1080/17517575.2024.2541199

Yang, X., Wang, X., Zhang, Q., Petzold, L., Wang, W. Y., & Zhao, X. (2023). Shadow alignment: The ease of subverting safely-aligned language models. arXiv:2310.02949. DOI: 10.48550/arXiv.2310.02949

Yuan, Z., Shang, Y., Zhou, Y., Dong, Z., Xue, Z., Wu, B., et al. (2024). QServe: W4A8KV4 quantization and system co-design for efficient LLM serving. arXiv:2405.04532. DOI: 10.48550/arXiv.2405.04532

Zerbib, O. D. (2019). The effect of pro-environmental preferences on bond prices: Evidence from

green bonds in the secondary market. Journal of Banking & Finance, 98, 39–60. DOI: 10.1016/j.jbankfin.2018.10.012

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

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. DOI: 10.1002/sres.3040

Zhang, J., Yang, X., & Appelbaum, D. (2015). Toward effective big data analysis in continuous auditing. Accounting Horizons, 29(2), 469–476. DOI: 10.2308/acch-51070

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

Zhu, M., & Gupta, S. (2018). To prune, or not to prune: Exploring the efficacy of pruning for model compression. arXiv:1710.01878. DOI: 10.48550/arXiv.1710.01878