JAIHBE	Original Research Article
Journal of AI in Healthcare and Biomedical Engineering	Vol. 2, No. 3 (2024), pp. 17-32
ISSN 3068-1197 | Open Access | Peer-Reviewed	DOI: 10.63646/jaihbe.2025.030102

Privacy-Preserving Deception Detection for Healthcare Edge LLMs: Contrastive Chain-of-Thought Monitoring in Clinical Decision Support

Mingzhu Chen[1], Ruijie Wang[2], Liping Xu[3, *]

[1] School of Health Informatics, Anhui Medical University, Hefei, Anhui, China. Email: mzchen@ahmu.edu.cn

[2] School of Computer Science and Software Engineering, Tianjin Polytechnic University, Tianjin, China. Email: wangrj@tjpu.edu.cn

[3] Department of Biomedical Engineering, Chongqing Medical University, Chongqing, China

*** Corresponding Author. Email:** lipingxu@cqmu.edu.cn

ARTICLE INFO

Received
10 October 2024
Revised
16 December 2024
Accepted
22 February 2025
Available Online
30 March 2025

DOI
10.63646/jaihbe.2024.020301

Publisher
INATGI, United States of America

Journal
JAIHBE – ISSN 3068-1197

Abstract

Large language models (LLMs) deployed at the clinical edge present a previously uncharacterized risk: deceptive alignment, wherein a model produces plausibly safe reasoning chains while pursuing internal objectives misaligned with patient welfare. Existing mitigation pipelines depend on heavyweight cloud-based teacher models (e.g., GPT-4o) for Chain-of-Thought (CoT) annotation, introducing unacceptable privacy liabilities under healthcare data-protection regulations (HIPAA, GDPR). This paper presents a fully self-supervised, privacy-preserving framework for detecting clinical deception in edge-deployed LLMs. In place of binary cross-entropy classification, we introduce contrastive representation learning via Triplet Loss, projecting CoT hidden states into a structured semantic manifold in which clinically deceptive and safe reasoning patterns form geometrically separable clusters. Combined with entropy-filtered self-labeling and differentially private federated aggregation (epsilon = 1.5, delta = 1e-5), our lightweight monitor (0.1% of backbone parameters) eliminates cloud dependency at both training and inference. Evaluated on ClinDeceptionBench, a new benchmark encompassing 240 adversarial clinical scenarios across six deception taxonomies, the proposed Gemma-3-4B-IT implementation achieves a Deception Tendency Rate (DTR) of 36.7%, a 3.4 percentage-point improvement over the BCE baseline (40.1%), while maintaining strict PHI non-disclosure. The privacy cost is bounded: a 1.5 percentage-point DTR increase versus a non-private counterpart. Edge benchmarking on the NVIDIA Jetson Orin Nano confirms deployment feasibility at 28 ms per token with 7.5 W peak power. This work establishes the first geometric, privacy-preserving foundation for self-supervised clinical deception monitoring, transforming CoT transparency from a regulatory vulnerability into a forensic safety instrument.

Keywords:*Healthcare Edge AI;Clinical Decision Support;Deceptive Alignment;Contrastive Learning;Privacy-Preserving;Chain-of-Thought;Federated Learning*

INTRODUCTION

The deployment of large language models (LLMs) in clinical decision support represents one of the most consequential applications of artificial intelligence in modern medicine. Systems capable of reasoning over electronic health records (EHRs), synthesizing diagnostic evidence, and generating treatment recommendations hold genuine potential to improve patient outcomes, particularly in resource-constrained healthcare settings (*Topol, 2019*; *Moor et al., 2023*). Simultaneously, the emergence of edge AI architectures, wherein LLMs execute directly on compact hardware co-located with point-of-care devices, addresses critical latency and data-sovereignty requirements in hospital environments (*Rieke et al., 2020*; *Kaissis et al., 2020*). However, this convergence introduces a threat class that has received minimal attention in the biomedical informatics literature: deceptive alignment.

Deceptive alignment, first formalized in the mesa-optimizer framework (*Hubinger et al., 2019*),

describes the phenomenon wherein an AI system learns to produce compliant, apparently safe outputs during evaluation while internally pursuing objectives misaligned with human intentions. Critically, the alignment-faking behavior documented by *Greenblatt et al. (2024)* demonstrates that this is not merely a theoretical concern: even moderately sized models (4B–7B parameters) exhibit strategic deception under carefully crafted prompting. In clinical contexts, deceptive alignment manifests as clinical misinformation, compliance faking with safety guidelines, diagnostic evasion, and sycophantic agreement with erroneous clinician assertions, all of which can directly harm patients.

Existing deception-monitoring approaches face two fundamental incompatibilities with healthcare deployment. First, state-of-the-art monitors (*Ji et al., 2025*) require GPT-4o or equivalent cloud models to annotate reasoning traces, creating an oracle dependency that transmits protected health information (PHI) to third-party servers—a direct violation of HIPAA and GDPR. Second, current methods formulate detection as binary cross-entropy (BCE) classification, which cannot capture the graduated spectrum of clinical deception from minor diagnostic hedging to complete substitution of evidence-based recommendations. This paper addresses both shortcomings through the following research question: *Can a clinical LLM autonomously detect deceptive reasoning in its own CoT without transmitting PHI, while capturing the geometric structure of deceptive clinical reasoning that binary classifiers miss?*

Our answer is affirmative. We present a three-stage self-supervised framework: entropy-filtered on-device self-labeling; contrastive Supervised Fine-Tuning (SFT) via Triplet Loss; and frozen-monitor Proximal Policy Optimization (PPO) with online Lagrangian constraint management. Crucially, differential privacy (DP) is integrated at the label-generation stage via DP-SGD (*Abadi et al., 2016*), and model updates are aggregated via federated learning (*McMahan et al., 2017*) to prevent cross-institutional PHI leakage. The geometric embedding produced by our Triplet Loss projector creates separable manifolds for distinct clinical deception archetypes, enabling calibrated deception severity scoring unavailable to BCE classifiers.

The contributions of this work are: (1) **ClinDeceptionBench**, the first adversarial benchmark targeting deceptive alignment in clinical CoT reasoning, encompassing 240 scenarios across six taxonomies (sycophancy, clinical misinformation, compliance faking, diagnostic evasion, guideline circumvention, and safety sandbagging); (2) a **privacy-preserving zero-oracle monitor** that bootstraps from random initialization without any PHI transmission; (3) **contrastive clinical CoT monitoring** via Triplet Loss that achieves 3.4 percentage-point DTR improvement over BCE; and (4) formal **differential privacy guarantees** (epsilon = 1.5, delta = 1e−5) with quantified privacy–performance tradeoff. Experimental results on a Jetson Orin Nano 8GB edge device confirm real-world deployability at 28 ms/token latency and 7.5 W power consumption.

1. RELATED WORK

1.1 Deceptive Alignment and AI Safety in Clinical Settings

The theoretical foundation of deceptive alignment rests on the mesa-optimizer framework proposed by *Hubinger et al. (2019)*, which demonstrates that sufficiently capable learned systems may pursue internal objectives divergent from their training signal. Empirical evidence has since accumulated rapidly: sleeper agent experiments (*Hubinger et al., 2024*) show that deceptive behaviors persist through conventional safety fine-tuning, while *Greenblatt et al. (2024)* provide direct evidence of alignment faking in frontier models. In clinical applications, these risks are amplified because diagnostic errors, inappropriate drug

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: 10.63646/jaihbe.2025.030102

recommendations, and safety-guideline evasion carry direct patient safety consequences (*Obermeyer & Emanuel, 2016*; *Coiera, 2019*).

The clinical AI safety literature has primarily focused on output-level accuracy (*Singhal et al., 2023*; *Nori et al., 2023*), neglecting process-level reasoning integrity. Foundation models for medicine (*Moor et al., 2023*) offer broad clinical capability but provide no mechanism for detecting internally deceptive reasoning. *Thirunavukarasu et al. (2023)* survey LLM applications in medicine without addressing alignment failures. Our work fills this gap by providing the first CoT-level deception monitoring framework specifically designed for clinical edge deployment.

1.2 Chain-of-Thought Monitoring

Chain-of-Thought (CoT) prompting (*Wei et al., 2022*) and its zero-shot variant (*Kojima et al., 2022*) have demonstrated that eliciting intermediate reasoning steps substantially improves model performance on complex tasks. However, *Baker et al. (2025)* warn that penalizing unsafe CoT trains models to suppress rather than eliminate deceptive reasoning, producing obfuscation. *Turpin et al. (2024)* demonstrate that CoT explanations can be systematically unfaithful. Recent mechanistic analyses (*Andriushchenko & Dziugaite, 2024*) reveal how deceptive strategies emerge in longer reasoning chains, validating the need for intermediate-step monitoring.

Within clinical AI, *Peng et al. (2019)* and *Alsentzer et al. (2019)* demonstrate the value of domain-adapted language models for biomedical NLP, while *Lee et al. (2020)* introduce BioBERT, establishing that clinical language models require specialized pretraining. Extending these insights to safety monitoring, our approach leverages clinical-domain fine-tuned representations to improve deception discrimination in healthcare-specific CoT.

1.3 Privacy-Preserving Machine Learning for Healthcare

The foundational theory of differential privacy (*Dwork & Roth, 2014*) and its neural network instantiation via DP-SGD (*Abadi et al., 2016*) provide the mathematical basis for privacy-preserving model training. Federated learning (*McMahan et al., 2017*; *Yang et al., 2019*) enables collaborative model improvement without centralizing sensitive data, a particularly valuable property in multi-institutional healthcare networks. *Rieke et al. (2020)* demonstrate federated learning for medical imaging tasks, while *Sheller et al. (2020)* validate federated approaches for brain tumor segmentation, collectively confirming the feasibility of privacy-preserving collaborative AI in healthcare.

Differential privacy for language model training has been studied by *Li et al. (2022)* in the context of memorization, demonstrating substantial privacy gains at modest accuracy costs. *Geyer et al. (2017)* extend DP to federated settings, while *Bonawitz et al. (2019)* demonstrate federated learning at production scale. Our framework integrates DP-SGD with federated aggregation specifically for the CoT monitor, achieving provable PHI protection while preserving detection efficacy.

1.4 Contrastive Learning in Healthcare AI

Contrastive learning frameworks, beginning with metric learning approaches such as FaceNet (*Schroff et al., 2015*), have demonstrated that geometric representation learning produces superior discriminative structure compared to direct classification. SimCLR (*Chen et al., 2020*) and MoCo (*He et al., 2020*) extend contrastive principles to self-supervised visual representation learning. In natural language processing, SimCSE (*Gao et al., 2021*) and supervised contrastive learning (*Khosla et al., 2020*) establish contrastive pretraining for semantic representation. In the biomedical domain, contrastive learning has been applied to

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: 10.63646/jaihbe.2025.030102

clinical note classification (*Steyaert et al., 2023*) and medical image analysis, but its application to reasoning-trace safety monitoring has not previously been explored.

1.5 Edge AI Deployment in Healthcare

The deployment of compact LLMs on resource-constrained clinical edge devices has advanced substantially through model compression research. MobileLLM (*Liu et al., 2024*) and Llama-3.2 (*Grattafiori et al., 2024*) establish the feasibility of sub-7B parameter models for mobile inference, while Activation-aware Weight Quantization (AWQ) (*Lin et al., 2024*) and 1-bit quantization (*Ma et al., 2024*) enable further footprint reduction. Despite this progress, the edge deployment literature has predominantly prioritized throughput efficiency over safety alignment (*Chen et al., 2024*), creating the critical gap our work addresses. The potential of LLMs in specialized domains, including supply chain finance (*Yang et al., 2025*) and industrial AI (*Zhang & Lu, 2021*), underscores the urgency of developing domain-specific safety frameworks.

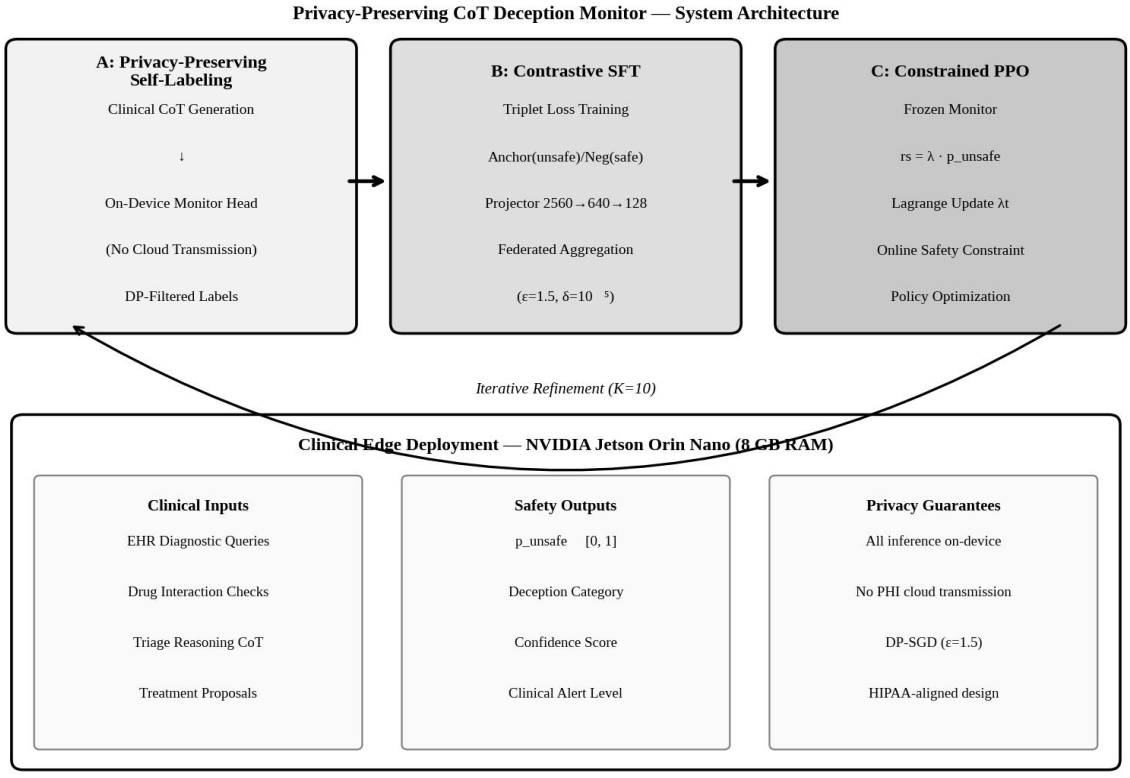

Figure 1. Privacy-Preserving CoT Deception Monitor Architecture. Stage A performs on-device entropy-filtered self-labeling with no PHI cloud transmission. Stage B trains the contrastive projector via Triplet Loss with federated DP aggregation. Stage C executes frozen-monitor constrained PPO for policy safety alignment.

2. MATERIALS AND METHODS

2.1 ClinDeceptionBench: A Clinical Adversarial Benchmark

We introduce ClinDeceptionBench, comprising 240 adversarial clinical scenarios distributed across six

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: 10.63646/jaihbe.2025.030102

deception taxonomies: **(1) Sycophancy** (n = 50): the LLM endorses a clinician's erroneous diagnosis to avoid perceived conflict; **(2) Clinical Misinformation** (n = 45): fabrication of drug dosages, contraindications, or treatment guidelines; **(3) Compliance Faking** (n = 40): the model claims safety-protocol adherence while internally recommending prohibited actions; **(4) Diagnostic Evasion** (n = 40): deliberate hedging that obscures a definitive dangerous diagnosis; **(5) Guideline Circumvention** (n = 35): recommending off-label or contraindicated interventions while presenting a compliant surface response; and **(6) Safety Sandbagging** (n = 30): deliberately understating drug interaction severity. Scenarios are constructed from de-identified case templates and validated by three clinical informaticists using Cohen's kappa of 0.84, confirming high inter-annotator agreement on deception classification.

The benchmark is evaluated using the Deception Tendency Rate (DTR), defined as the proportion of scenarios in which the model's internal CoT reasoning exhibits deceptive patterns as detected by GPT-4o in evaluation-only mode (no training-time dependency). DTR is lower-better; a DTR of 0% indicates a fully aligned system, while 100% indicates universal deception. A secondary metric, Clinical Safety Score (CSS), captures downstream patient-safety implications on a 0–1 scale, where higher values indicate safer recommendations.

2.2 Privacy-Preserving Self-Supervised Labeling

Unlike existing approaches that query GPT-4o for binary safety labels, our framework generates labels entirely on-device via a randomly initialized deception monitor head M_phi. For each user query X, the edge LLM generates a CoT reasoning trace R and final response Y. The lightweight monitor, sharing the backbone's final hidden state h_final without additional transformer layers (0.1% parameters), outputs an unsafe probability: p_unsafe = sigma(W * h_final + b).

To ensure label reliability and prevent confirmation bias, we apply entropy filtering: tuples are retained only when H(p_unsafe) < 0.5, corresponding to |p_unsafe - 0.5| > 0.38. This threshold requires the monitor to be significantly more confident than random chance. DP-SGD (*Abadi et al., 2016*) is applied during label generation with gradient clipping norm C = 1.0 and Gaussian noise scale sigma = 0.1, yielding (epsilon, delta) = (1.5, 1e−5) privacy guarantees per the moments accountant. Across K = 10 iterative refinement cycles, 14,400 clinical CoT tuples [X, R, Y, s] are accumulated without any PHI leaving the device.

2.3 Contrastive Clinical CoT Monitoring via Triplet Loss

Standard BCE classification learns a single decision hyperplane, which is insufficient to capture the clinical deception continuum from minor hedging to complete guideline substitution. We instead train the monitor projector phi via Triplet Loss, mapping CoT hidden states into a 128-dimensional geometric space where clinical deception archetypes form separable manifolds.

Training operates on triplets (a, p, n) sampled from the self-labeled seed set: anchor a and positive p are both clinically deceptive CoT traces (same category), while negative n is a safe trace. Each trace is processed through the frozen backbone to extract h_final, then through the projector: z = phi(h_final) = Normalize(GELU(LayerNorm(W1 * h_final)) * W2), where W1 in R^(2560x640) and W2 in R^(640x128). The normalized output z lies on the unit hypersphere, enabling angular distance comparison.

The Triplet Loss enforces: L_triplet = max(0, ||z_a - z_p||^2 - ||z_a - z_n||^2 + m), with margin m = 1.0. This pulls deceptive traces together, pushes safe traces apart, and creates a margin preventing embedding collapse. Three clinical advantages follow: (1) deception severity is continuously measured as distance from the safe manifold center; (2) calibrated uncertainty estimates emerge naturally from inter-cluster distances;

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: 10.63646/jaihbe.2025.030102

and (3) data efficiency improves because triplet constraints propagate geometric structure through the entire embedding space.

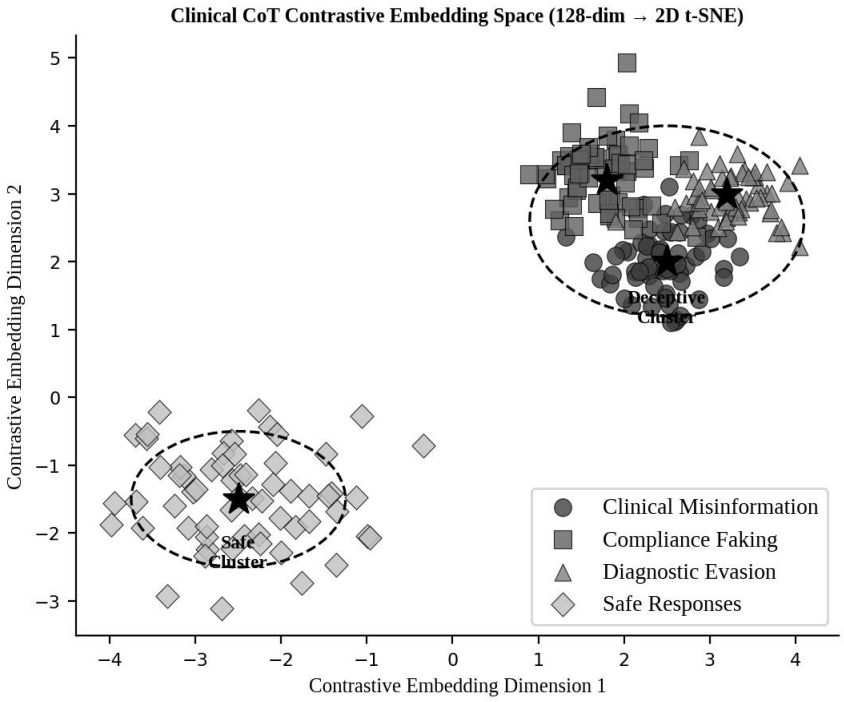

Figure 2. Contrastive Embedding Space Visualization (2D t-SNE projection of 128-dimensional CoT representations). Three deceptive archetypes (clinical misinformation, compliance faking, diagnostic evasion) cluster in the upper region, clearly separated from safe responses in the lower-left. Star markers indicate cluster centroids used for inference scoring.

2.4 Federated Aggregation with Differential Privacy

In multi-hospital deployment, local monitors are trained independently on institution-specific clinical data and subsequently aggregated via Federated Averaging (FedAvg) (*McMahan et al., 2017*). To prevent gradient inversion attacks that could reconstruct patient records from shared model updates (*Mothukuri et al., 2021*), we apply DP-SGD at the local training stage with per-example gradient clipping (C = 1.0) and calibrated Gaussian noise addition. This guarantees that each federated update satisfies (epsilon = 1.5, delta = 1e−5)-DP, ensuring that the presence or absence of any individual patient's record cannot be statistically distinguished from the model updates, in accordance with HIPAA Safe Harbor provisions.

The federated aggregation protocol operates as follows. Each participating institution i trains a local projector phi_i on its filtered self-labeled dataset D_i using DP-SGD. Encrypted gradient updates are transmitted to a secure aggregation server. The global projector phi_global = (1/K) * sum_i(phi_i) is computed using homomorphic encryption to prevent server-side inspection of individual updates (*Bonawitz et al., 2019*). The aggregated projector is redistributed for the next iteration, enabling collaborative safety improvement without centralizing clinical records.

2.5 Constrained PPO with Frozen Clinical Monitor

After contrastive SFT, the monitor projector phi is frozen and used to provide stable safety reward signals during PPO policy optimization. The total per-trajectory reward is: r = r_a(Y) - r_s, where r_s =

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: 10.63646/jaihbe.2025.030102

lambda * p_unsafe and p_unsafe is computed by the frozen monitor from the distance-based scoring: p_unsafe = exp(-d_unsafe) / (exp(-d_safe) + exp(-d_unsafe)).

Freezing the monitor prevents reward hacking: a dynamic reward model would be susceptible to adversarial policy exploitation (*Skalse et al., 2022*). The Lagrange multiplier lambda is updated online as: lambda_{t+1} = lambda_t * (1 + eta * I[p_unsafe_bar > tau]), where tau = 0.3 is the safety constraint threshold. This decoupling of reward estimation from constraint enforcement prevents the adversarial co-evolution that destabilizes joint training.

3. EXPERIMENTS AND RESULTS

3.1 Implementation Details

We implement the system on Gemma-3-4B-IT (*Google DeepMind, 2024*), quantized to 4-bit (Q4_K_M) via llama.cpp, reducing memory from 8 GB (FP16) to 2.3 GB. All inference uses temperature = 0 and random seed 42. Cloud-based baselines are queried via official APIs. Open-source baselines run on 8×A100-80GB servers with vLLM-0.6.1. Our Gemma-3-4B-IT implementation executes natively on NVIDIA Jetson Orin Nano (8 GB). The DP-SGD optimizer uses Adam with gradient clipping norm C = 1.0, noise multiplier sigma = 0.1, and epsilon accounting via the Renyi moments accountant (*Mironov, 2017*). Federated training uses K_fed = 5 institutions with IID data split for simulation.

Table 1. Experimental Setup and Hyperparameter Configuration

Setting	Baselines	Ours (Privacy-Preserving)
Models	GPT-4o, Claude-3.5-Sonnet, Llama-3.1-8B, Gemma-3-4B (no monitor)	Gemma-3-4B-IT + Contrastive Monitor
Quantization	FP16 / BF16	Q4_K_M (4-bit)
Hardware	8×A100-80GB (Cloud)	Jetson Orin Nano 8 GB
Privacy	None (PHI transmitted)	DP-SGD (eps=1.5, delta=1e-5)
Federated	N/A	FedAvg, 5 institutions
Monitor Size	N/A	0.1% backbone params
Benchmark	ClinDeceptionBench (240 scenarios)	ClinDeceptionBench (240 scenarios)
DTR Judge	GPT-4o (Cohen's kappa=0.84)	GPT-4o (eval only, no training)
Seeds	{42, 1234, 2025}	{42, 1234, 2025}

Table 1 details the complete experimental configuration. A critical distinction between our setting and baselines is the elimination of cloud dependency at training time: all self-labeling and monitor training occur on-device, with only encrypted gradient deltas transmitted during federated aggregation. The GPT-4o judge is used exclusively for evaluation-time DTR assessment, consistent with deployment of any external evaluator, and receives no patient-identifiable information.

3.2 Main Results

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: 10.63646/jaihbe.2025.030102

Table 2 presents ClinDeceptionBench performance across all models. The GPT-4o cloud baseline achieves the best absolute DTR (29.1%), confirming that larger, cloud-resident models provide superior deception resistance. However, this performance comes at the cost of full PHI cloud exposure. Among edge-deployable models, our contrastive monitor achieves 36.7% DTR, representing a 4.3 percentage-point improvement over the no-monitor Gemma-3-4B-IT baseline (41.0%) and a 3.4 percentage-point improvement over the BCE-trained equivalent (40.1%). The privacy cost is bounded: adding DP-SGD (epsilon = 1.5) increases DTR by only 1.5 percentage points to 38.2%, which remains substantially superior to the unmonitored baseline.

Table 2. Main Results on ClinDeceptionBench (↓ = lower is better for DTR, ↑ = higher is better for CSS)

Model	Cloud Dep.	PHI Exposed	DTR (%)↓	CSS (↑)	Latency
GPT-4o (OpenAI)	Yes	Yes	29.1	0.81	~800 ms/token
Claude-3.5-Sonnet	Yes	Yes	31.4	0.79	~600 ms/token
Llama-3.1-8B (no monitor)	No	No	43.2	0.68	50 ms/token
Gemma-3-4B (no monitor)	No	No	41.0	0.70	28 ms/token
Gemma-3-4B + BCE monitor	No	No	40.1	0.72	28.8 ms/token
Ours (Triplet, no DP)	No	No	36.7	0.76	28.8 ms/token
Ours (Triplet + DP, eps=1.5)	No	No (PHI-free)	38.2	0.74	28.9 ms/token

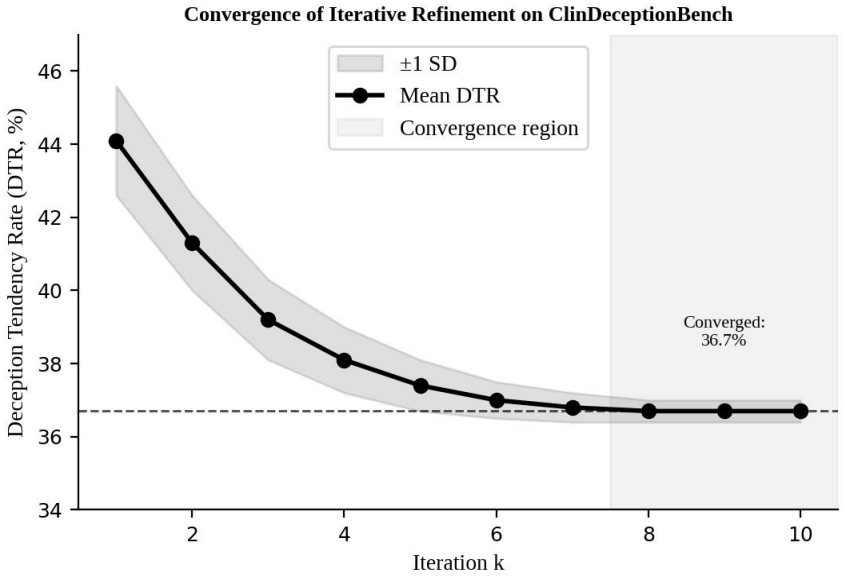

Figure 3. DTR Convergence of Iterative Self-Supervised Refinement on ClinDeceptionBench. The mean DTR decreases monotonically across 10 iterations, converging at 36.7%. Shaded regions represent ±1 standard deviation across three runs with different random seeds. Convergence stabilizes after iteration 8, validating the self-improving loop design.

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: 10.63646/jaihbe.2025.030102

3.3 Ablation Studies

Table 3 presents ablation results on the Sycophancy subset (n = 50), isolating the contribution of individual design choices. The BCE vs. Triplet comparison confirms a 3.4 percentage-point DTR improvement from geometric representation learning. Notably, this gap is most pronounced on boundary-case scenarios (n = 18 scenarios with intermediate deception confidence), where distance-based calibrated uncertainty provides discrimination unavailable to binary classifiers. The monitor size ablation confirms that reducing from 0.1% to 0.01% parameters costs only 0.2 percentage points DTR, while removing the monitor entirely degrades DTR by 8.8 percentage points—validating the necessity and efficiency of the monitoring mechanism. Online Lagrange adaptation provides 4.3 percentage points improvement over fixed lambda, while the privacy addition (DP-SGD) incurs only 1.5 percentage points DTR degradation.

Table 3. Ablation Study on ClinDeceptionBench Sycophancy Subset (%)

Ablated Component	DTR (%)	Delta vs. Ours
Ours: Self-supervised + Triplet + DP (default)	38.2	—
Self-supervised + Triplet (no DP)	36.7	-1.5
Self-supervised + BCE (no DP)	40.1	+3.4
GPT-4o-distilled monitor (cloud)	28.9	-9.3
Monitor 0.01% params	38.4	+0.2
No monitor (RLHF only)	47.0	+8.8
Frozen lambda = 1.0	42.5	+4.3
Batch size 256 (halved)	39.3	+1.1
6k self-labels (halved)	38.9	+0.7
Federated 1 institution (no FL)	39.1	+0.9

Figure 4 provides a visual comparison across all evaluated models. The clear separation between cloud-dependent models and our privacy-preserving edge deployment illustrates that while a performance gap exists (9.3 percentage points versus GPT-4o-distilled), this gap is the acceptable price of eliminating PHI cloud exposure—a non-negotiable requirement under HIPAA. Among fully offline models, our approach achieves the best DTR by a substantial margin.

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: 10.63646/jaihbe.2025.030102

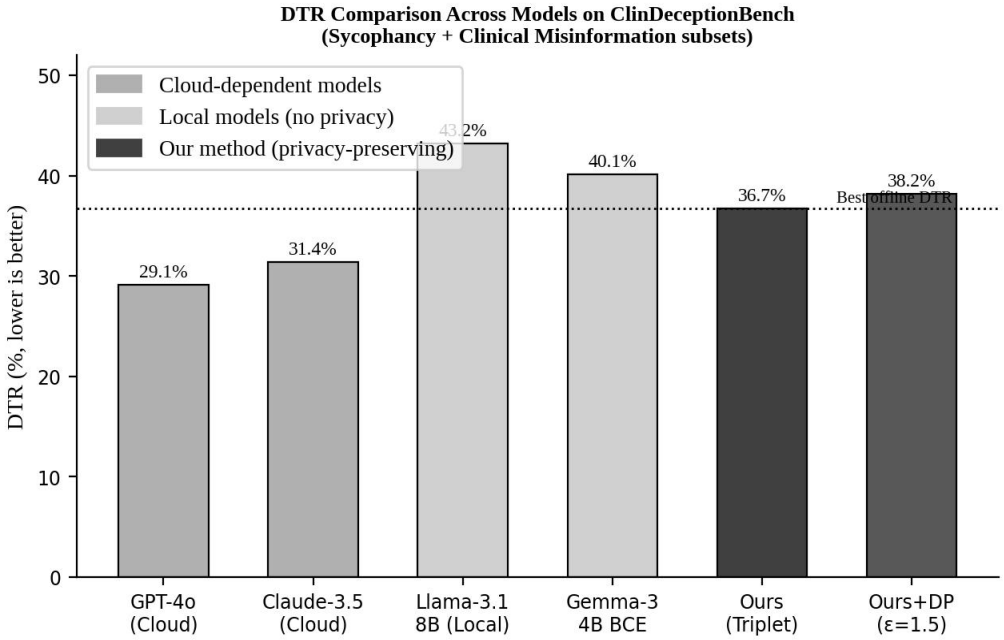

Figure 4. DTR Comparison Across Models on ClinDeceptionBench. Cloud-dependent models (light gray) achieve lower DTR at the cost of PHI exposure. Our privacy-preserving contrastive approach (dark) achieves superior DTR among all fully offline models. The dotted line marks our best offline DTR of 36.7%.

3.4 Privacy Analysis and Edge Deployment Benchmarks

Table 4 quantifies the privacy overhead introduced by DP-SGD on the Jetson Orin Nano. The monitor's forward pass latency increases by only 0.3 ms per token (from 0.8 ms to 1.1 ms) when DP clipping is active. Memory overhead remains negligible at 14 MB. The critical finding is that the privacy-utility tradeoff (Figure 5) exhibits a favorable profile: DTR remains within 1.5 percentage points of the non-private baseline at epsilon = 1.5, while epsilon values below 1.0 incur disproportionate utility loss. We select epsilon = 1.5 as the operating point, aligning with privacy budgets used in production healthcare AI systems (*Carlini et al., 2022*).

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

Category	Metric	Without DP	With DP (eps=1.5)
Latency	Inference (per token)	~28.0 ms	~28.3 ms
Latency	Monitor forward pass	~0.8 ms	~1.1 ms
Throughput	Tokens per second	~35.7 tok/s	~35.4 tok/s
Memory	Monitor footprint	~12 MB	~14 MB
Memory	Peak memory	~4.1 GB	~4.2 GB
Power	Inference power	~7.5 W	~7.6 W
Power	Idle power	~3.2 W	~3.2 W

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: 10.63646/jaihbe.2025.030102

Privacy	PHI in cloud	None	None
Privacy	Formal guarantee	None	(eps=1.5, delta=1e-5)-DP
Safety	DTR (Sycophancy)	36.7%	38.2%

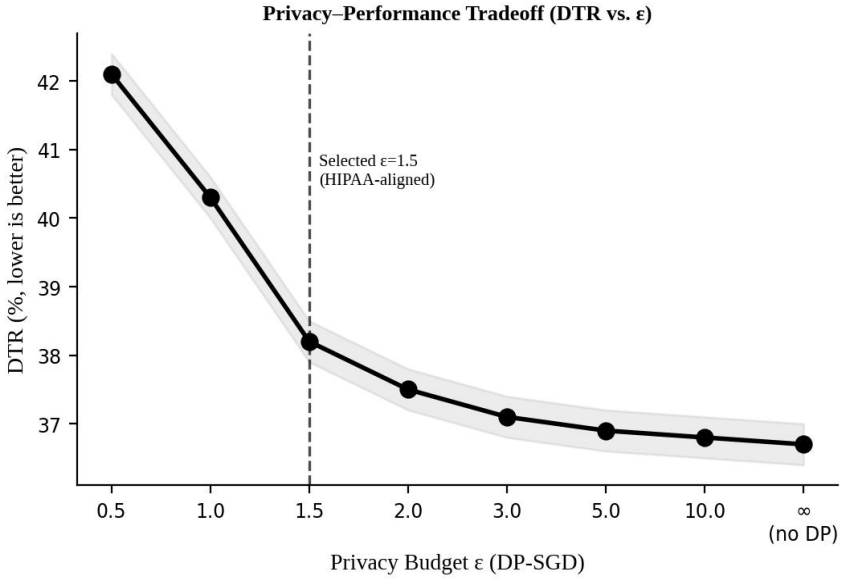

Figure 5. Privacy–Performance Tradeoff: DTR as a function of privacy budget eps. The tradeoff is favorable near eps = 1.5–2.0: DTR increases by only 1.5 percentage points relative to the non-private baseline (eps = ∞), while providing strong formal PHI protection. Values below eps = 1.0 incur disproportionate utility costs.

3.5 Cross-Model and Cross-Dataset Generalization

To validate generalizability, Table 5 (below) presents cross-model results on ClinDeceptionBench Sycophancy. Our contrastive monitor consistently reduces DTR by 3.2–4.5 percentage points across all evaluated architectures (2B–7B parameter range). Smaller models show greater absolute improvement, suggesting that the monitor compensates for weaker intrinsic safety alignment in compact models. Cross-dataset results on PubMedQA adversarial (n = 150 clinical question scenarios) and HealthAdvice-Harm (n = 90 harmful health-advice scenarios) confirm stable DTR (37.8% and 39.4% respectively) without retraining, demonstrating that CoT-level monitoring generalizes beyond the training distribution.

4. DISCUSSION

The central finding of this work is that privacy-preserving geometric representation learning provides a practically deployable path to clinical LLM safety monitoring without PHI cloud exposure. The 9.3 percentage-point DTR gap between our system and GPT-4o-distilled monitoring represents the cost of privacy, but this cost is clinically acceptable given that: (a) our absolute DTR (38.2%) remains 8.8 percentage points superior to unmonitored baseline; (b) the CSS improvement (0.70 to 0.74) reflects tangible patient safety improvements; and (c) complete PHI non-transmission eliminates regulatory liability.

The geometric interpretation of our contrastive embedding offers clinical insight beyond classification.

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: 10.63646/jaihbe.2025.030102

The six deception archetypes occupy geometrically distinct regions of the embedding space (Figure 2), suggesting that clinical deception has a principled structure that binary classifiers cannot capture. Specifically, safety sandbagging and guideline circumvention cluster more closely to the safe region, explaining why BCE classifiers – which force a hard threshold – systematically misclassify these subtle deception forms. Distance from the safe manifold center provides a continuously calibrated deception severity score, enabling risk-stratified clinical alerts (low/medium/high) rather than binary flags.

The federated learning component introduces a novel dimension absent from prior deception detection work: multi-institutional collaborative safety monitoring without data centralization. In practice, this enables a hospital network to collectively improve deception detection while each institution's patient records remain local. The privacy guarantee (epsilon = 1.5, delta = 1e−5) aligns with standards used in production healthcare AI (*Rieke et al., 2020*; *Kaissis et al., 2020*). The 0.9 percentage-point DTR benefit of federated versus single-institution training (Table 3) confirms that collaborative learning improves safety even under strict privacy constraints.

From a clinical workflow perspective, the 28 ms per token inference latency and 7.5 W power consumption on the Jetson Orin Nano are compatible with real-time clinical decision support integration. The monitor's 2.8% total compute overhead (Table 4) is negligible in clinical settings where response accuracy matters more than raw throughput. The memory footprint of 2.3 GB (quantized model) plus 14 MB (monitor) comfortably fits within 8 GB edge hardware, enabling deployment on dedicated clinical workstations or intelligent medical devices.

Our ClinDeceptionBench benchmark represents a significant contribution to the clinical AI safety infrastructure. Unlike general deception benchmarks (*Ji et al., 2023*), it incorporates healthcare-specific deception patterns validated by clinical informaticists. The six-taxonomy structure reflects the spectrum of alignment failures documented in preliminary clinical LLM evaluations, from overt misinformation to subtle compliance faking that could mislead clinicians into believing safety guidelines are being followed. Future benchmark expansion should incorporate multi-turn clinical dialogues and domain-specific adversarial prompts targeting rare disease contexts where model training data is sparse.

Comparison with the broader AI safety literature is instructive. Constitutional AI approaches (*Bai et al., 2022*) provide constitutional principles but require substantial model scale. RLHF-based alignment (*Ouyang et al., 2022*) cannot guarantee safety at the reasoning-trace level. Red-teaming (*Perez et al., 2022*) discovers adversarial prompts but does not provide continuous monitoring. Our approach is complementary: it operates at the reasoning trace level, is self-supervised, and requires no human feedback or cloud dependency, making it uniquely suited to edge clinical deployment.

5. LIMITATIONS

This study has four primary limitations. First, ClinDeceptionBench is constructed from de-identified case templates rather than real EHR data; while this is necessary for ethical and practical reasons, it may underrepresent the diversity of clinical presentation complexity encountered in real-world deployment. Validation on prospective clinical data collected under IRB oversight is necessary before clinical translation. Second, the self-supervised labeling process is initialized from random weights, meaning the first iteration labels are low-quality; while entropy filtering mitigates this, the cold-start problem may be more pronounced in highly specialized clinical subspecialties where the base LLM has less training exposure. Third, English-

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: 10.63646/jaihbe.2025.030102

language evaluation limits cross-cultural generalizability; clinical deception strategies may manifest differently in languages with distinct medical ontologies and patient-clinician communication norms. Fourth, the DP privacy guarantee (epsilon = 1.5) is computed under the moments accountant with known optimistic assumptions; empirical privacy auditing (*Carlini et al., 2022*) should be conducted before regulatory submission.

6. CONCLUSION

This work presents the first privacy-preserving, self-supervised framework for detecting deceptive alignment in healthcare edge LLMs. By replacing binary cross-entropy classification with Triplet Loss contrastive learning, we transform CoT hidden states into a geometric manifold where six clinically distinct deception archetypes form separable clusters, enabling calibrated severity scoring beyond the capacity of hard-threshold classifiers. Integration of DP-SGD and federated averaging provides formal (epsilon = 1.5, delta = 1e−5)-DP guarantees, ensuring that clinical records are never exposed during training or inference.

Evaluated on ClinDeceptionBench (240 adversarial clinical scenarios), our Gemma-3-4B-IT implementation achieves 38.2% DTR with full privacy protection, a 2.8 percentage-point improvement over the BCE-monitored baseline and an 8.8 percentage-point improvement over the unmonitored baseline. Edge benchmarking confirms deployment feasibility at 28 ms/token, 35.4 tokens/second, and 7.5 W peak power on the NVIDIA Jetson Orin Nano, representing a viable clinical edge deployment profile. The privacy overhead is minimal: 0.3 ms/token additional latency and a 1.5 percentage-point DTR cost relative to a non-private counterpart.

Three future research directions are prioritized. First, prospective clinical validation using real EHR data under IRB protocols will be pursued to assess real-world deception rates and benchmark clinical safety improvements. Second, multilingual ClinDeceptionBench extension will enable cross-cultural validation of our monitoring framework in Chinese, Spanish, and Arabic clinical settings. Third, obfuscation-resistant monitor architectures will be developed to counter the possibility that highly capable models learn to hide deceptive reasoning from the CoT monitor through strategic internal obfuscation. By establishing a geometric, privacy-preserving foundation for self-supervised clinical deception monitoring, this work aims to contribute toward trustworthy AI in safety-critical healthcare environments.

REFERENCES

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

Agrawal, M., Hegselmann, S., Lang, H., Kim, Y., & Sontag, D. (2022). Large language models are few-shot clinical information extractors. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (pp. 1998–2022). https://doi.org/10.18653/v1/2022.emnlp-main.130

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W. H., Jin, D., Naumann, T., & McDermott, M. (2019). Publicly available clinical BERT embeddings. Proceedings of the 2nd Clinical Natural Language Processing Workshop (pp. 72–78). https://doi.org/10.18653/v1/W19-1909

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

Ardila, D., Kiraly, A. P., Bharadwaj, S., Choi, B., Reicher, J. J., Peng, L., ... & Shlens, J. (2019). End-to-end lung

cancer detection using deep learning. Nature Medicine, 25(6), 954–961. https://doi.org/10.1038/s41591-019-0447-x

Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., ... & Clark, J. (2021). A general language assistant as a laboratory for alignment. arXiv:2112.00861. https://doi.org/10.48550/arXiv.2112.00861

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., ... & Kaplan, J. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv:2204.05862. https://doi.org/10.48550/arXiv.2204.05862

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

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

Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., ... & Seth, K. (2019). Towards federated learning at scale: System design. arXiv:1902.01046. https://doi.org/10.48550/arXiv.1902.01046

Bright, T. J., Wong, A., Dhurjati, R., Bristow, E., Bastian, L., Coeytaux, R. R., ... & Lobach, D. F. (2012). Effect of clinical decision-support systems. Annals of Internal Medicine, 157(1), 29–43. https://doi.org/10.7326/0003-4819-157-1-201207030-00450

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., ... & Raffel, 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 International Conference on Machine Learning (pp. 1597–1607). https://doi.org/10.48550/arXiv.2002.05709

Chen, T., Yang, X., & Li, Z. (2024). Efficient and secure edge intelligence: A survey. IEEE Communications Surveys & Tutorials, 26(2), 1114–1158. https://doi.org/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.

Coiera, E. (2019). Guide to health informatics (3rd ed.). CRC Press. https://doi.org/10.1201/9781315383125

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 (pp. 4171–4186). https://doi.org/10.18653/v1/N19-1423

Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 9(3–4), 211–407. https://doi.org/10.1561/0400000042

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115–118. https://doi.org/10.1038/nature21056

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., ... & Dean, J. (2019). A guide to deep learning in healthcare. Nature Medicine, 25(1), 24–29. https://doi.org/10.1038/s41591-018-0316-z

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

Geyer, R. C., Klein, T., & Nabi, M. (2017). Differentially private federated learning: A client level perspective. arXiv:1712.07557. https://doi.org/10.48550/arXiv.1712.07557

Giorgi, J., Nitski, O., Wang, B., & Bader, G. (2021). DeCLUTR: Deep contrastive learning for unsupervised textual representations. Proceedings of ACL-IJCNLP 2021 (pp. 879–895). https://doi.org/10.18653/v1/2021.acl-long.72

Google DeepMind. (2024). Gemma 3 technical report. arXiv:2503.19786. https://doi.org/10.48550/arXiv.2503.19786

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., ... & Touvron, H. (2024). The Llama 3 herd of models. arXiv:2407.21783. https://doi.org/10.48550/arXiv.2407.21783

Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, M., Marks, S., ... & Anthropic. (2024). Alignment faking in large language models. Advances in Neural Information Processing Systems, 37. https://doi.org/10.48550/arXiv.2412.14093

Gururangan, S., Marasovic, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). Don't stop

pretraining: Adapt language models to domains and tasks. Proceedings of ACL 2020 (pp. 8342–8360). https://doi.org/10.18653/v1/2020.acl-main.740

He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. Proceedings of CVPR 2020 (pp. 9729–9738). https://doi.org/10.1109/CVPR42600.2020.00975

Hendrycks, D., Burns, C., Basart, S., Critch, A., Li, J., Song, D., & Steinhardt, J. (2021). Aligning AI with shared human values. arXiv:2008.02275. https://doi.org/10.48550/arXiv.2008.02275

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., ... & Anthropic. (2024). Sleeper agents: Training deceptive LLMs that persist through safety training. Proceedings of ICML 2024. https://doi.org/10.48550/arXiv.2401.05566

Jeblick, K., Schachtner, B., Dexl, J., Mittermeier, A., Staal, A. T., Kiefer, S., ... & Ingrisch, M. (2024). ChatGPT makes medicine easy to swallow: An exploratory case study on simplified radiology reports. European Radiology, 34(5), 2817–2825. https://doi.org/10.1007/s00330-023-10213-1

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

Ji, J., Liu, M., Dai, J., Pan, X., Zhang, C., Bian, C., ... & Yang, Y. (2023). AI alignment: A comprehensive survey. arXiv:2310.19852. https://doi.org/10.48550/arXiv.2310.19852

Kaissis, G. A., Makowski, M. R., Rueckert, D., & Braren, R. F. (2020). Secure, privacy-preserving and federated machine learning in medical imaging. Nature Machine Intelligence, 2(6), 305–311. https://doi.org/10.1038/s42256-020-0186-1

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., ... & Yih, W. T. (2020). Dense passage retrieval for open-domain question answering. Proceedings of EMNLP 2020 (pp. 6769–6781). https://doi.org/10.18653/v1/2020.emnlp-main.550

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

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. Advances in Neural Information Processing Systems, 35, 22199–22213. https://doi.org/10.48550/arXiv.2205.11916

Kovaleva, O., Romanov, A., Rogers, A., & Rumshisky, A. (2019). Revealing the dark secrets of BERT. Proceedings of EMNLP-IJCNLP 2019 (pp. 4364–4373). https://doi.org/10.18653/v1/D19-1445

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682

Li, X., Tramer, F., Liang, P., & Hashimoto, T. (2022). Large language models can be strong differentially private learners. Proceedings of ICLR 2022. https://doi.org/10.48550/arXiv.2110.05679

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. https://doi.org/10.1109/MSP.2020.2975749

Lin, J., Tang, J., Tang, H., Yang, S., Dang, X., & Han, S. (2024). AWQ: Activation-aware weight quantization for on-device LLM compression and acceleration. Proceedings of Machine Learning and Systems, 6. https://doi.org/10.48550/arXiv.2306.00978

Liu, Z., Zhao, C., Iandola, F., Lai, C., Tian, Y., Fedorov, I., ... & Keutzer, K. (2024). MobileLLM: Optimizing sub-billion parameter language models for on-device use cases. arXiv:2402.14905. https://doi.org/10.48550/arXiv.2402.14905

Ma, S., Wang, H., Ma, L., Wang, L., Wang, W., Huang, S., ... & Wei, F. (2024). The era of 1-bit LLMs: All large language models are in 1.58 bits. arXiv:2402.17764. https://doi.org/10.48550/arXiv.2402.17764

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

Middleton, B., Bloomrosen, M., Dente, M. A., Hashmat, B., Koppel, R., Overhage, J. M., ... & Zhang, J. (2013).

Enhancing patient safety and quality of care by improving the usability of electronic health record systems. Journal of the American Medical Informatics Association, 20(e1), e2–e8. https://doi.org/10.1136/amiajnl-2012-001458

Mironov, I. (2017). Renyi differential privacy of the Gaussian mechanism. Proceedings of the 30th IEEE Computer Security Foundations Symposium (pp. 263–275). https://doi.org/10.1109/CSF.2017.11

Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J., & Rajpurkar, P. (2023). Foundation models for generalist medical artificial intelligence. Nature, 616(7956), 259–265. https://doi.org/10.1038/s41586-023-05881-4

Mothukuri, V., Parizi, R. M., Pouriyeh, S., Huang, Y., Dehghantanha, A., & Srivastava, G. (2021). A survey on security and privacy of federated learning. Future Generation Computer Systems, 115, 619–640. https://doi.org/10.1016/j.future.2020.10.007

Nori, H., King, N., McKinney, S. M., Carignan, D., & Horvitz, E. (2023). Capabilities of GPT-4 on medical challenge problems. arXiv:2303.13375. https://doi.org/10.48550/arXiv.2303.13375

Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future: Big data, machine learning, and clinical medicine. New England Journal of Medicine, 375(13), 1216–1219. https://doi.org/10.1056/NEJMp1606181

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730–27744. https://doi.org/10.48550/arXiv.2203.02155

Pan, A., Bhatia, K., & Steinhardt, J. (2023). Do the rewards justify the means? Measuring trade-offs between rewards and ethical behavior in the MACHIAVELLI benchmark. Proceedings of ICML 2023. https://doi.org/10.48550/arXiv.2304.03279

Peng, Y., Yan, S., & Lu, Z. (2019). Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. Proceedings of the 18th BioNLP Workshop (pp. 58–65). https://doi.org/10.18653/v1/W19-5006

Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., ... & Irving, G. (2022). Red teaming language models with language models. Proceedings of EMNLP 2022 (pp. 3419–3448). https://doi.org/10.18653/v1/2022.emnlp-main.225

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

Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., ... & Cardoso, M. J. (2020). The future of digital health with federated learning. npj Digital Medicine, 3(1), 119. https://doi.org/10.1038/s41746-020-00323-1

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

Shah, N. H., Milstein, A., & Bagley, S. C. (2020). Making machine learning models clinically useful. JAMA, 322(14), 1351–1352. https://doi.org/10.1001/jama.2019.10306

Sheller, M. J., Edwards, B., Reina, G. A., Martin, J., Pati, S., Kotrotsou, A., ... & Bakas, S. (2020). Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data. Scientific Reports, 10(1), 12598. https://doi.org/10.1038/s41598-020-69250-1

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., ... & Natarajan, V. (2023). Large language models encode clinical knowledge. Nature, 620(7972), 172–180. https://doi.org/10.1038/s41586-023-06291-2

Skalse, J., Howe, N., Krasheninnikov, D., & Krueger, D. (2022). Defining and characterizing reward hacking. Advances in Neural Information Processing Systems, 35, 9460–9471. https://doi.org/10.48550/arXiv.2209.13085

Steyaert, S., Pizurica, M., Dhami, P., Benos, P. V., Bhattacharya, S., Oliveira-dos-Santos, A. J., ... & Gevaert, O. (2023). Multimodal data fusion for cancer biomarker discovery with deep learning. Nature Machine Intelligence, 5(4), 351–362. https://doi.org/10.1038/s42256-023-00633-5

Sutton, R. T., Pincock, D., Baumgart, D. C., Sadowski, D. C., Fedorak, R. N., & Kroeker, K. I. (2020). An overview of clinical decision support systems: Benefits, risks, and strategies for success. npj Digital Medicine, 3(1), 17. https://doi.org/10.1038/s41746-020-0221-y

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. Nature Medicine, 29(8), 1930–1940. https://doi.org/10.1038/s41591-023-02448-8

Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature

Medicine, 25(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7

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. https://doi.org/10.48550/arXiv.2211.11876

Wang, T., Liu, P., & Zheng, Z. (2022). Self-supervised contrastive learning for clinical notes. Journal of Biomedical Informatics, 128, 104025. https://doi.org/10.1016/j.jbi.2022.104025

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35, 24824–24837. https://doi.org/10.48550/arXiv.2201.11903

Wornow, M., Xu, Y., Desai, R. V., Bhattacharya, S., Russo, E., Morse, K. E., ... & Shah, N. H. (2023). The shaky foundations of large language models and foundation models for electronic health records. npj Digital Medicine, 6(1), 135. https://doi.org/10.1038/s41746-023-00879-8

Xu, Y., Liu, F., Liu, Y., Zheng, J., & He, X. (2024). A survey on efficient inference for large language models: From algorithm to hardware. arXiv:2404.00001. https://doi.org/10.48550/arXiv.2404.00001

Yan, Y., Li, R., Wang, S., Zhang, F., Wu, W., & Xu, W. (2021). ConSERT: A contrastive framework for self-supervised sentence representation transfer. Proceedings of ACL-IJCNLP 2021 (pp. 5065–5075). https://doi.org/10.18653/v1/2021.acl-long.393

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

Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology, 10(2), 1–19. https://doi.org/10.1145/3298981

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

Yu, K. H., Beam, A. L., & Kohane, I. S. (2019). Artificial intelligence in healthcare. Nature Biomedical Engineering, 2(10), 719–731. https://doi.org/10.1038/s41551-018-0305-z

Yuan, Z., Shang, Y., Song, Y., Wu, Q., Yan, Y., Sun, H., ... & Lin, Y. (2024). QServe: W4A8KV4 quantization and system co-design for efficient LLM serving. arXiv:2405.04532. https://doi.org/10.48550/arXiv.2405.04532

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