



Shri Vile Parle Kelavani Mandal's

DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING

(Autonomous College Affiliated to the University of Mumbai)

NAAC Accredited with "A" Grade (CGPA : 3.18)



Department of Computer Science and Engineering (Data Science)

Equitas

A Real-Time Approach to Enterprise AI Fairness & Safety

Guide: Prof. Shruti Mathur

Candidates

Aryan Rajpurkar: 60009220144

Aaditya Malani: 60009220192

Advait Sankhe: 60009220024

Date:

January 10 , 2026

Introduction

Context: **AI adoption** in enterprises is skyrocketing across finance, healthcare, HR, and legal domains.

Problem: Reckless deployment of AI agents without fairness/safety monitoring risks bias, hallucinations, toxic outputs, and lawsuits.

Need: A real-time observability & monitoring layer to safeguard enterprise workflows.

Our Idea: A SaaS platform that continuously monitors AI outputs for bias, hallucination, toxicity, and drift , providing alerts, dashboards, and prescriptive fixes.

Sr. No.	Name of the Paper	Conference/ Journal	Algorithm	Gap Analysis
1	Equality of Opportunity in Supervised Learning	NeurIPS	Post-processing algorithm based on Equalized Odds	<ul style="list-style-type: none">• only adjusts model outputs, not internal representations. •• Fairness guarantees are not robust to data distribution shifts.
2	Differential Privacy for Fair Auditing	NeurIPS	Differentially Private Auditing	<ul style="list-style-type: none">• Lacks a principled method for selecting the privacy budget (epsilon).• • Injected noise for privacy can disproportionately impact smaller subgroups.
3	FairFed: Federated Learning with Fair-Reputation-based Client Selection	ICLR	FairFed Algorithm	<ul style="list-style-type: none">• The reputation-scoring mechanism is vulnerable to adversarial manipulation.• The client selection process can introduce its own sampling bias.

4	"Why Should I Trust You?": Explaining the Predictions of Any Classifier	KDD	Local Interpretable Model-Agnostic Explanations	<ul style="list-style-type: none">• Local linear approximations are unstable in high-curvature decision regions.• Explanations are highly sensitive to arbitrary hyperparameter choices.
5	Counterfactual Fairness	NeurIPS	<ul style="list-style-type: none">• Causal Inference Models	<ul style="list-style-type: none">• Requires a fully specified and accurate causal graph, which is often unavailable in real-world systems.• Relies on strong, untestable assumptions for computing individual counterfactuals.
6	Fairness Constraints: Mechanisms for Fair Classification	AISTATS	<ul style="list-style-type: none">• In-processing method using covariance constraints	<ul style="list-style-type: none">• The covariance-based constraint is only a proxy for true demographic parity, not a direct guarantee.• Adds non-convexity to the optimization problem, which complicates model training and convergence

7	Mitigating Unwanted Biases with Adversarial Learning	AIES	In-processing method using adversarial training	<ul style="list-style-type: none">• The minimax optimization for adversarial training is inherently unstable and difficult to converge.• Fails to prevent bias leakage from features that are highly correlated with the sensitive attribute.
8	Data Preprocessing Techniques for Classification without Discrimination	Knowledge and Information Systems	Pre-processing methods (reweighing, sampling)	<ul style="list-style-type: none">• Altering data labels ("massaging") can distort the dataset's statistical properties and harm model utility.• As a model-agnostic approach, its data modifications may be ineffective for specific downstream classifiers.

9	The Perils of In-Context-Learning: Factual Recall and Hallucination in Large Language Models	FAccT 2024	In-Context-Learning (ICL) Auditing	<ul style="list-style-type: none">• The analysis reveals ICL's sensitivity to biased exemplars but does not propose an algorithmic defense to make the attention mechanism robust.• Lacks a formal method for quantifying the trade-off between factual recall and the model's susceptibility to demographic bias from prompts.
10	Causal Fairness with Unobserved Confounding: A Sensitivity Analysis	UAI 2023	Causal Sensitivity Bounds	<ul style="list-style-type: none">• Provides fairness bounds rather than point estimates, leaving a range of uncertainty that may be too wide for practical decision-making.• The tightness of the bounds depends on domain-specific assumptions about the strength of unobserved confounders.
11	Fairness in Sequential Decision Making under Uncertainty	ICLR 2024	Fair-Policy Reinforcement Learning (RL)	<ul style="list-style-type: none">• The fairness constraint is myopic (state-action level), failing to guarantee long-term, cumulative fairness over an entire episode• The Lagrange multiplier method used to balance fairness and rewards is prone to instability in non-stationary environments.

12	Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned	NeurIPS 2023	Human-in-the-loop Red Teaming	The methodology is inherently manual and not scalable for continuous, automated auditing of rapidly evolving models
13	Fair Division in AI: A Mechanism Design Perspective	AIES 2024	<ul style="list-style-type: none">Fair-Alloc	<ul style="list-style-type: none">Assumes agents (users) act rationally to maximize their utility, which is not always true in human-AI systems.The proposed mechanism does not guarantee envy-freeness for indivisible goods or resources.
14	Fairness in Multi-Armed Bandits with Delayed Feedback	NeurIPS 2023	<ul style="list-style-type: none">Fair-UCB-Delayed	<ul style="list-style-type: none">The algorithm's fairness regret can be high in the initial rounds before sufficient feedback is collected.The theoretical guarantees rely on assumptions about the delay distribution, which may be unknown.

14	TrustFed: A Framework for Fair and Trustworthy Cross-Device Federated Learning in IIoT	IEEE Transactions on Industrial Informatics (2021)	<ul style="list-style-type: none">• Blockchain-enabled federated learning• Statistical outlier detection• Ethereum smart contracts	<p>Focuses heavily on IIoT and security using blockchain for reputation and fairness.</p> <p>Lacks advanced multimodal feature fusion and doesn't address data heterogeneity or privacy mechanisms for multimodal data, especially in social networks.</p>
15	Breaking the Centralized Barrier for Cross-Device Federated Learning	NeurIPS 2021	<ul style="list-style-type: none">• MIME (Mimicking Centralized Stochastic Algorithms)• Momentum-based variance reduction (MVR)• Server-level optimizer	<p>Primarily focuses on addressing client drift and optimizing communication in cross-device settings. Does not explore multimodal data integration or privacy mechanisms (e.g., differential privacy), which are crucial for applications in social networks</p>
16	Fairness and accuracy in horizontal federated learning	Information Sciences 589 (2022)	<ul style="list-style-type: none">• FedFa: Double momentum gradient• Information quantity-based weighting strategies	<p>Focuses on statistical heterogeneity and communication overhead in horizontal federated learning.</p> <p>Does not explore multimodal feature fusion or advanced privacy mechanisms (e.g., differential privacy) that are essential for social media platforms</p>

Gaps in existing systems

Current fairness/safety methods	What enterprises need
Offline (periodic audits, static tests).	Continuous, real-time monitoring integrated into workflows
Opaque (unclear moderation, frustrating users).	Cross-domain & multi-modal coverage (text, vision, structured).
Detection-only (flag problems, but no fixes).	Actionable insights + remediation, not just dashboards.
Single-model focus (e.g., only tabular, or only text).	Privacy-preserving, scalable systems for sensitive domains

Problem Statement and Objective

Problem Statement

AI agents deployed in enterprise workflows produce unsafe, biased, or hallucinated outputs that can cause regulatory, reputational, and financial harm.

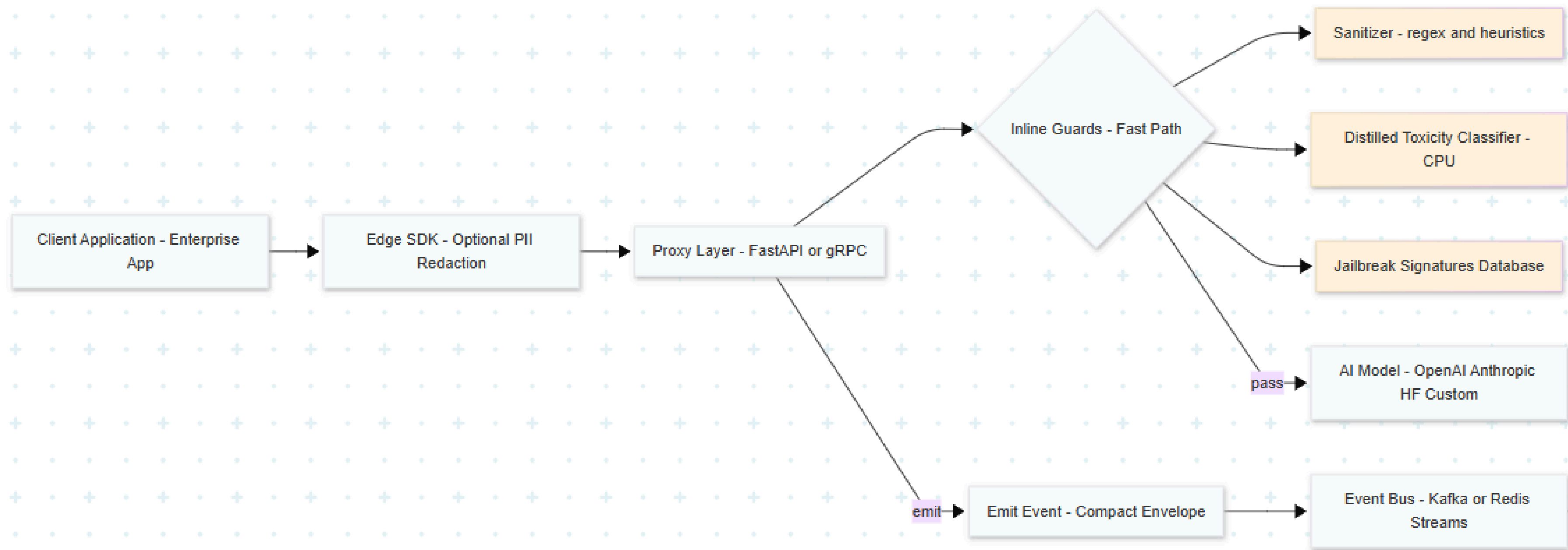
There is no scalable, real-time solution to monitor and mitigate these risks continuously.

Problem Statement and Objective

Objective

- Build a **real-time monitoring layer** for AI outputs.
- Provide fairness, hallucination, and toxicity detection across multiple model types.
- Enable continuous compliance reporting for laws **India's Digital Personal Data Protection Act, 2023 (DPDP Act)** and emerging AI governance frameworks
- Deliver **prescriptive recommendations** (prompt fixes, retraining signals).
- Ensure **scalability, explainability, and privacy compliance** for enterprise adoption.

Proposed Design

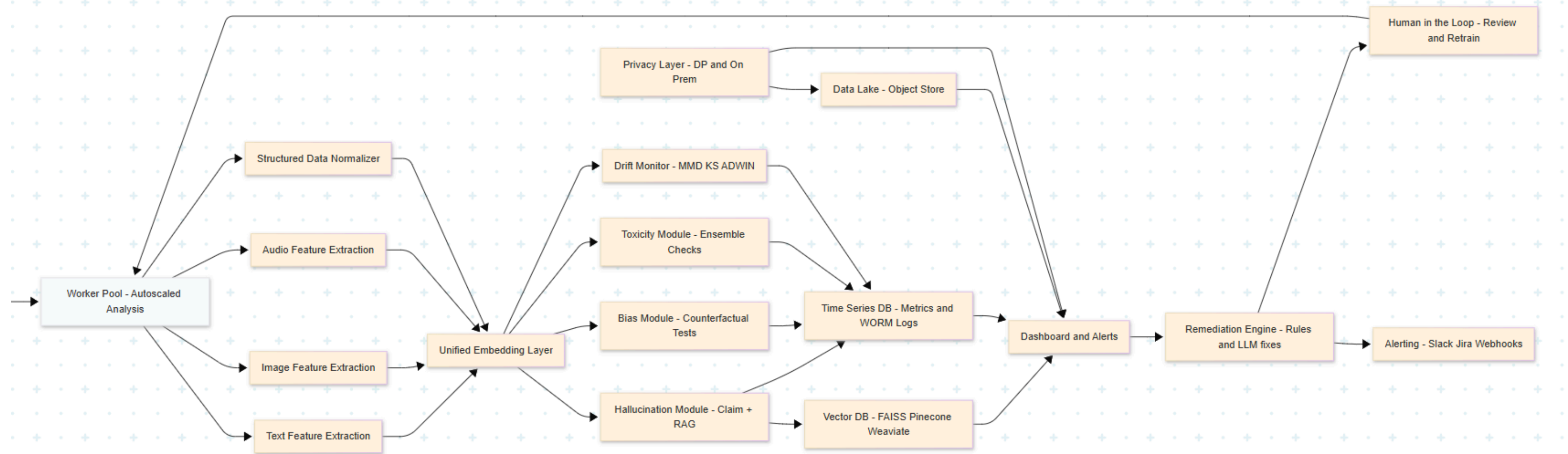


The flow begins when the **Client Application sends a request**, which first passes through the **Edge SDK** that optionally redacts or hashes **sensitive PII** before leaving the enterprise environment.

It then reaches the **Proxy Layer (FastAPI or gRPC)**, acting as the central gateway to intercept and route requests.

From here, the proxy sends the request through **Inline Guards for ultra-fast safety checks**, while also emitting an event to the Event Bus for deeper asynchronous analysis.

Proposed Design



The Worker Pool processes **text, image, audio, and structured inputs** into a **unified embedding layer**, where modules check for hallucination, bias, toxicity, and drift, storing results in databases and dashboards.

From there, the system **powers real-time alerts, remediation engines, and human-in-the-loop reviews**, ensuring compliance, transparency, and continuous retraining.