Madhava Gaikwad

Senior Technical Leader – AI Infrastructure & Distributed Systems

18 years building scalable Network Security, AI/ML systems and leading engineering teams

gaikwad.madhav@gmail.com Bengaluru, India linkedin.com/in/safeai github.com/krimler Google Scholar

Technical Skills

Languages: C/C++, Python

AI/ML Infrastructure: PyTorch, Ray, Triton Inference Server, LangChain, LlamaIndex, Langfuse, HuggingFace Transformers, MLflow, Weights & Biases, FastAPI, RAG architectures, RLHF production systems

Distributed Systems: Kubernetes, Docker, Redis, Microservices, Distributed consensus, gRPC, HTTP/2, Event-driven architectures

Security & Privacy: SSL/TLS, PKI, Cryptography, Differential Privacy, Identity & IAM, OAuth2, Zero Trust, Container Security, Policy-as-Code

Cloud & Deployment: Azure ML, Azure AKS, Helm, Vector DBs (FAISS, Pinecone)

Observability & Reliability: Prometheus, Grafana, OpenTelemetry, SLI/SLO frameworks, Performance profiling, Incident automation, Chaos Engineering

Professional Experience

Microsoft Azure Networking Feb 2021–Present

Senior Engineer – AI Infrastructure & Technical Leadership, Bangalore

- **Served as Engineering Manager (2023-2025)** leading 15-person team building AI reliability systems; promoted 4 engineers to senior roles, defined technical strategy for LLM-based SRE tools. Currently IC role focused on AI infrastructure research and systems building.
- **Built and led delivery of RMA-nator**, a forecasting platform for proactive hardware management; designed probabilistic triage algorithms reducing datacenter technician toil by 35%.
- **Architected Net Copilot**, an LLM-based recommender system serving 200+ engineers across four teams; achieved 30% faster Mean Time to Mitigate through ranked remediation and actionability gating.
- **Developed production RLHF frameworks** (NPO, Opal) translating reinforcement-learning feedback into measurable reliability gains in live Azure datacenters.
- **Designed Alignment-Ops infrastructure** with RL-based alert ranking and adaptive thresholding, reducing alert fatigue by 40%.
- **Built CoreSec**, distributed consensus system for network RCA achieving 99.98% availability with formally verified safety guarantees.
- **Deployed adaptive WAF rules** using production multi-armed bandits, cutting false positives by 20%.

Cisco Systems Jun 2012–Feb 2021

Security & Infrastructure, Bangalore

Technical Leader Feb 2019–Feb 2021

- **Led architecture and delivery** of $80M Web Security Gateway across three global teams

(18 engineers); designed cloud adaptation strategy and enterprise integration patterns for Fortune 500 customers.

- **Drove cross-organization security strategy** and architecture reviews; collaborated with product, cloud, and research teams to align features with evolving enterprise threat models.

Engineering Manager Dec 2016–Feb 2019

- **Managed team of 8 developers** delivering core proxy backend components on schedule; promoted 2 to SDE II and 2 to SDE III.
- **Designed and shipped 10+ major security features** including on-box DLP, HTTP/2 enablement, unified PKI, and proxy performance optimizations for Fortune 500 customers.

Senior Engineer Jun 2012–Dec 2016

- **Engineered TLS performance optimizations** in FreeBSD TCP/IP stack and multi-threaded TLS application layer, achieving 1.2× throughput and 2× decryption performance through pipeline redesign.
- **Implemented privacy-preserving telemetry framework** using differential privacy guarantees, ensuring GDPR compliance at production scale.

Symantec Jul 2010–Jun 2012

Software Engineer – Security Products, Pune

- Developed IPv6 stack and integration test framework for Linux-based SMTP security service; resolved critical memory-management issues.

Persistent Systems Jan 2010–Jul 2010

Senior Software Engineer, Pune

- Built extension modules for enterprise SMTP service in C++ and Java.

Tata Consultancy Services Dec 2006–Jan 2010

Software Engineer, Mumbai

- Developed AAA (Authentication, Authorization, Accounting) module for telecom switch in C and Java.

Research & Production Innovation

Security & Privacy Systems

- **AlignDP:** Differential privacy for LLM feedback telemetry with rare-event protection — deployed in Azure production pipelines (NeurIPS 2025 Lock-LLM Workshop)
- **Adaptive WAF:** Production multi-armed bandits for Web Application Firewall rule tuning — deployed across Azure regions, reduced false positives by 20% through RL-based optimization PhilPapers
- **AVEC:** Adaptive verifiable edge control for local LLMs with per-query differential privacy and delegation auditing arXiv:2509.10561
- **Privacy-Preserving Telemetry:** Differential privacy framework for enterprise security products ensuring GDPR compliance at production scale (deployed at Cisco 2016-2019)

AI/ML & Reinforcement Learning Systems

- **NPO:** Structured RLHF framework with meta-alignment — production deployment in Azure datacenters, measurable policy correction arXiv:2507.21131
- **Opal:** Operator algebra framework unifying RLHF methods with canonical schema arXiv:2509.11298 (submitted to ICLR 2026)

- **Murphy's Laws of AI Alignment:** Lower bounds on RLHF under feedback misspecification arXiv:2509.05381
- **Two-Knob Control Theory:** Low-dimensional projection for supervisory controls; Control Sufficiency Index for LLM decoding strategies [paper]
- **Memomind:** Criticality-stratified learning for AI agent memory with cache adjudication policies — deployed in agent orchestration [paper]

SRE, Reliability & Observability Systems

- **Structure Reduces Chaos:** Schema-driven operational data management at datacenter scale — deployed system processing 15,000+ tickets, reduced MTTR by 50%, $2.1M annual savings through automated evidence collection and STAR/ECO extraction (submitted to SIGMOD 2026)
- **RMA-nator:** Forecast-guided link management at hyperscale — forecasting and confidence-fusion platform reducing datacenter toil by 35% TechRxiv
- **Alignment Metric:** Feedback-driven drift detection measuring operator-automation agreement; demonstrated MTTR correlation (18→27→12 min) across multiple datacenters (manuscript in preparation)
- **ANSC:** Probabilistic capacity health scoring combining failure prediction with resource constraints — deployed across 400+ datacenters, 60 regions arXiv:2508.16119
- **CoreSec:** Distributed consensus for network Root-Cause Analysis with formal safety guarantees — 2+ years production, 99.98% availability, mechanized verification [paper]

Education

Bachelor of Engineering (Information Technology) — Shivaji University Kolhapur, 2006

Awards & Recognition

- Microsoft "Pinnacle: Copilot of the Year!" Award (2025)
- Microsoft Azure "Quality of Service Excellence" Award (2024)
- Microsoft Azure "Quality of Service Excellence" Award (2023)
- Microsoft Azure "Leadership in AI Systems Engineering" Award (2023)
- Microsoft Azure "Quality of Service Excellence" Award (2022)
- Cisco Technical Leadership Award for TLS and HTTP optimization (2016)