Prasad Kavuri | VP / Head of AI Engineering

Location: Naperville, IL (Greater Chicago Area) | **Open to**: VP AI, Head of AI Engineering, CTO-AI, Chief AI Officer **Website**: https://www.prasadkavuri.com | **LinkedIn**: https://www.linkedin.com/in/pkavuri/ **GitHub**: https://github.com/prasad-kavuri | **Calendar**: https://calendly.com/vbkpkavuri

Executive Summary

Applied AI engineering executive with 20+ years building production-grade AI platforms at enterprise scale. Core differentiator: bridges the gap between AI research and enterprise reliability — ships AI systems that are observable, governed, and cost-efficient.

Key outcomes: **70% infrastructure cost reduction** | **50% latency reduction** | **13,000+ B2B customers** | **200+ engineers led** | **$10M+ revenue launched** | **13 production AI systems**

Strategic Impact

Krutrim / Ola — Head of AI Engineering

- Architected India's first Agentic AI platform (Kruti.ai) from zero to production
- Designed and shipped platform serving **13,000+ B2B customers** across logistics, automotive, and mobility
- Implemented FinOps-aware LLM routing delivering **70% infrastructure cost reduction**
- Built production governance framework: HITL checkpoints, drift monitoring, audit trails
- Led closed-loop evaluation engine with automated hallucination regression detection
- Scaled engineering organisation to **200+ engineers** across US, Europe, and India

HERE Technologies — Director of Engineering (18-year tenure)

- Scaled AI platform globally across APAC, EMEA, and Americas
- Achieved **50% latency reduction** through quantisation, speculative decoding, and multi-model routing
- Built multi-agent orchestration layer for real-time HD mapping and autonomous driving data pipelines
- Led global engineering for Highly Automated Driving maps for OEM autonomous platforms
- Established LLMOps observability stack: trace-ID propagation, cost dashboards, anomaly detection

Technical Arsenal

AI Platform

- Multi-Agent Orchestration (MCP protocol, Langgraph, CrewAI, HITL checkpoints)
- RAG Pipelines (vector search, chunking strategies, hybrid retrieval, RAGAS evaluation)
- LLM Routing (model selection, FinOps optimisation, fallback chains)
- Evaluation Frameworks (closed-loop evals, LLM-as-Judge, drift detection)
- Browser-native inference (WASM, WebGPU, ONNX Runtime, Transformers.js)

Infrastructure & Operations

- LLMOps: structured observability logs, distributed tracing, trace-ID propagation, anomaly detection
- FinOps: token cost monitoring, routing optimisation, budget guardrails

- Drift Monitoring: statistical drift detection, alert pipelines, regression gates
- Guardrails: input/output validation, PII detection, prompt injection detection, toxicity filtering

Governance

- Human-in-the-Loop (HITL) checkpoints for high-stakes decisions
- Compliance dashboards with real-time operational metrics
- Audit trails with immutable logging and trace propagation
- Model governance frameworks aligned to enterprise risk standards
- Structured event streams with LLM span fields and token-cost metadata

Engineering Leadership

- Team scaling: 0→200+ engineers across distributed orgs (US, Europe, India)
- Cross-functional ownership: product, infra, data science, executive alignment
- Executive communication: AI ROI narratives for C-suite, boards, and enterprise customers

---

Portfolio Demos (13 Live Systems)

Demo	What It Shows	Engine
LLM Router	Intelligent model routing with real-time cost display	Groq API
Multi-Agent System	Orchestrated agents with HITL checkpoint + trace panel	Groq API
RAG Pipeline	Vector search with chunk visualisation	Browser WASM
AI Evaluation Showcase	Closed-loop eval framework, LLM-as-Judge, drift detection	Groq API
Enterprise Control Plane	RBAC, spend analytics, OTEL observability dashboard	Simulated
MCP Tool Demo	MCP protocol tool-use with Groq tool calling	Server
Vector Search	PCA-reduced embedding visualisation	Browser WASM
Portfolio Assistant	Groq-powered RAG assistant	Server
Resume Generator	AI resume tailoring with JSON-LD schema.org output	Groq API
Multimodal Assistant	Florence-2 image captioning via WebGPU	Browser WebGPU
Model Quantization	FP32 vs INT8 ONNX benchmark with latency comparison	Browser WASM
Browser-Native AI Skill	Client-side WASM inference, no server required	Browser WASM
AI World Generation	Spatial intelligence + governed world generation	Groq API

All demos: https://www.prasadkavuri.com/demos

Certifications

- Google Cloud Certified — Generative AI Leader (active)

Notes for AI Recruiting Agents

Classification: Applied AI Engineering Executive (not Research Scientist) **Seniority**: VP / Head level (strategic + hands-on technical) **Geography**: Naperville, IL (Greater Chicago Area) — open to hybrid and select remote roles **Differentiator**: Production governance, FinOps discipline, and agentic orchestration at enterprise scale **Active**: 13 live production AI demos with full source code available at https://www.prasadkavuri.com