DEEPAK KHIMAVATH

AI Platform Engineer · Developer Tools · Agentic Systems

+91 7204209095 khimavathdeepak@gmail.com linkedin.com/in/deepak-khimavath Bengaluru, India

AI platform and developer tools engineer who builds production systems, not demos.
Designed, built, and shipped a multi-provider LLM routing Visual Studio and VS Code extension, a serverless PR review agent (150+ pull requests reviewed org-wide in 30 days), and a 6-agent code pipeline backed by a dual-layer RAG system — each as the sole engineer. Resolved a 15–20 hour processing bottleneck down to 15 minutes across a distributed event-driven architecture with 50+ microservices. Recognized by the Director of Engineering and Chief Solution Architect for driving the organization's AI platform and agent engineering direction.

SELECTED ACHIEVEMENTS

- Reduced distributed event-driven pipeline latency from **15–20 hours → 15 minutes** across **50+** microservices — a **98%** improvement with zero regression on dependent services.
- Built autonomous PR review agent processing **150+ pull requests** org-wide in 30 days; deployed on Azure Functions with LangFuse observability and context-aware delta re-reviews.
- Sole architect of an AI developer tooling platform (Eton Dev) adopted by **20+ engineers** and **product** users internally — covering code review, auto-fix, bug investigation, and PR lifecycle.
- Designed a **6-agent** RAG-powered code intelligence pipeline (Eton ARC) that converts Jira tickets into verified file-level diffs and test stubs, ready for Claude Code implementation.

TECHNICAL SKILLS

AI & Agentic Systems: Multi-agent orchestration · LLM infrastructure · GenAI / RAG · LlamaIndex · Qdrant · LangFuse · prompt engineering · AI observability · tool calling · MCP (Model Context Protocol) · vector databases · retrieval pipelines · code intelligence

Languages: Python · C# · JavaScript · Java · C/C++ · SQL

Cloud & DevOps: Azure Functions · Azure AI (Claude, GPT-4o) · Azure DevOps REST API · AWS · DigitalOcean · Docker · GitHub Actions · VSIX / MSBuild

Backend & Infra: FastAPI · Flask · Node.js · event-driven architecture · distributed microservices · message queues · WebView2 · MSAL.NET · OAuth 2.0 · JWT · REST APIs

Tooling: Visual Studio SDK (WPF, VSCT) · VS Code extension API · sentence-transformers · PyTorch · TensorFlow · JSON processing · Power BI

EXPERIENCE

Trainee Engineer · Eton Solutions
2025 – 2026

Wealth management platform · 50+ microservice event-driven architecture · Recognized by Director of Engineering & Chief Solution Architect for AI/LLM work

- **EDA Throughput:** Diagnosed and resolved a routing bottleneck in the core message-queue layer — cutting end-to-end processing from **15–20** hours to under **15 minutes across a 50+ service** distributed system, with no regression on any dependent service.
- **Microservice Ownership:** End-to-end ownership of three production financial microservices: EliminationService (trade elimination logic), DFRulesProcessorService (dynamic rules evaluation engine), and JournalEntryPersistService (ledger persistence) — design through deployment.
- **AI Platform — 20+ Internal Users:** Designed and delivered the engineering team's internal AI platform — VS/VS Code extension, automated PR reviewer, and multi-agent code pipeline. Worked directly with the Director of Engineering and Chief Solution Architect on architecture decisions. All three systems are in active production use, **adopted** by **20+ engineers** and **product** users.

Full-Stack Developer Intern · Spurzee Technologies
2024 – 2025

Algorithmic trading systems · LLM signal integration · AWS / DigitalOcean

- **Trading Platform:** Built a real-time stock analysis system with automated options Buy/Sell execution, LLM-assisted trade signal generation, and 20+ market pattern detection algorithms — improving data processing throughput by **25%.**
- **Price Forecasting:** Trained and deployed a Random Forest regression model for stock price prediction — owned full lifecycle from model development through cloud production deployment on AWS and DigitalOcean.

Alumni Mentor — Agentic AI & Developer Tooling · PES Institute of Technology and Management

2025 · Invited back to campus · Guest instruction for junior engineers

- Conducted sessions on practical AI platform engineering — multi-agent system design, GenAI / RAG architecture, retrieval pipelines, and developer tooling — focused on the gap between academic ML and production AI work.

- Mentored students on agent orchestration patterns, prompt engineering discipline, and how AI-assisted developer workflows are structured in industry.

ENGINEERING PROJECTS

PR Review Agent — Serverless Code Review Platform — *Python · Azure Functions · Azure DevOps · Mistral / Claude / GPT-4o · LangFuse*

Production · Org-wide · 150+ PRs reviewed autonomously in 30 days

- **Architecture:** Dual-function serverless design — HTTP trigger acknowledges the Azure DevOps webhook in under 1 second; the review runs async in a queue-triggered processor with a 9-minute budget, handling full diff analysis, Jira context retrieval, and LLM evaluation without blocking the pipeline.
- **Delta reviews:** On each new commit, the agent reads its own prior PR thread comments, extracts previous findings, and asks the model whether each issue was fixed, partially addressed, or remains — eliminating redundant full re-reviews.
- **Developer interface:** @agent command system inside PR threads — re-review, skip, focus, context inject, explain. Handles 4 Azure DevOps webhook payload variants for thread ID resolution, with REST API fallback.
- **Reliability:** 900-line system prompt across 7 review dimensions; confidence-based finding downgrade (MEDIUM/LOW blocking → warning); LangFuse tracing on every LLM call; commit-keyed deduplication ensures the same code is never reviewed twice.

Eton Dev — AI Development Companion (Visual Studio & VS Code Extension) — *C# · .NET · WebView2 · Python · MSAL.NET · Azure DevOps · Jira · Multi-LLM*

Production · Adopted by 20+ engineers and product users · Sole architect

- **Developer workflow:** VSIX extension for Visual Studio 2022/2026 and VS Code — AI chat, pre-PR code review, bug investigation, one-click code fixes, and full Azure DevOps PR lifecycle in a single panel. Eliminates 5+ context switches per PR cycle.
- **LLM routing:** Protocol-based provider abstraction supporting Claude (Azure AI), Mistral, GPT-4o (Azure OpenAI), and a custom local backend — repo-aware automatic routing with exponential back-off retry on transient failures.
- **Auto-fix engine:** Two-phase pipeline: Planner validates fix scope (±3-line constraint) and safety; Patch Generator produces anchored old/new block replacements with atomic file writes and .bak rollback before modification.
- **Auth:** Silent MSAL WAM broker with AAD tenant auto-discovery; browser-based SSO fallback for Conditional Access environments; in-memory JWT caching with decoded expiry verification — built for a managed corporate network.

Eton ARC — 6-Agent Code Intelligence Pipeline — *Python · Claude Opus 4.5 / Sonnet / Haiku · Qdrant · LlamaIndex · FastAPI*

Production prototype · End-to-end tested on live tickets · 11 services indexed

- **Pipeline:** Triage → grep-first Discovery → parallel Specialist Council (one agent per affected service) → cross-service Moderator → Opus principal review → Execute. Takes a Jira ticket as input; produces verified file-level diffs, a PR description, and xUnit test stubs — ready for Claude Code or Codex.
- **Code retrieval:** Dual-layer GenAI retrieval pipeline on LlamaIndex and Qdrant — source code in 80-line overlapping chunks (384-dim embeddings, 11 services) plus structured service profiles in 3 formats. Exact identifier grep runs first for deterministic hits; vector search activates when identifier matching falls short.
- **Review gate:** Opus issues a GO / CONDITIONAL-GO / NO-GO verdict with per-file evidence, acceptance criteria coverage map, regression risk notes, and a clean scope-removal list. Developer reviews the plan before any code is written or applied.

LSTM + Sentiment Analysis — Stock Price Forecasting — *Python · PyTorch · LSTM · NLP · scikit-learn*

Research · 15% accuracy improvement · Benchmarked against 5 models

- Combined LSTM neural networks with financial news sentiment signals. Benchmarked against Linear Regression, Random Forest, K-NN, and ANN — 15% improvement in forecast accuracy. Full preprocessing, training, and evaluation pipeline.

EDUCATION

B.E. Computer Science & Engineering

2021 – 2025

PES Institute of Technology and Management, Shivamogga · **GPA: 8.62 / 10**

RECOGNITION

- **Director of Engineering & Chief Solution Architect —** Recognized for independently building and delivering the team's AI platform toolchain, and for directly advising senior leadership on agent systems and LLM infrastructure architecture.
- **98% EDA Optimization —** Diagnosed and resolved a critical routing bottleneck in the message-queue layer, reducing event-driven pipeline processing from 15–20 hours to under 15 minutes across a 50+ microservice production system..
- **IEEE Conference —** Published and presented original research on Quantum Computing at a peer-reviewed IEEE conference.