VED CHITNIS

+91 97695 10044 ◇ Bengaluru, India
ved.chitnis@gmail.com ◇ linkedin.com/in/ved-chitnis ◇ vedwhat.com

SUMMARY

Software Engineer with 5+ years building scalable backend systems and AI-powered platforms at Cisco. Specialized in distributed systems, performance optimization, and LLM infrastructure. Strong track record in system architecture, reducing costs by 80%+ through intelligent design while improving performance by 90%+ and achieving 99.9% uptime at scale.

EXPERIENCE

Software Engineer Feb 2020 - Present
Cisco Systems - Third Party Software Compliance & Risk Management *Bengaluru, India*

Tech Stack: Python/FastAPI, Java/Spring Boot, React, PostgreSQL/pgvector, AWS (EKS, EC2, Bedrock), Azure OpenAI, Kubernetes, Terraform, Prometheus

- Architected and managed 7 AWS EKS production clusters from scratch orchestrating 100+ pods across 50+ microservices handling 1000+ QPS, implementing security hardening (RBAC, network policies), observability (Prometheus/Grafana), and CI/CD automation with 99.9% uptime.
- Built production LLM orchestration platform from scratch for performance optimization and customization, featuring custom agent executor with graph architecture, intelligent model routing (80% cost savings via query complexity analysis), and multi-tier inference (Azure OpenAI → Bedrock/Llama failover) achieving <800ms TTFT and 99.9% uptime.
- Engineered custom RAG infrastructure with hybrid retrieval (BM25 + semantic via PostgreSQL/pgvector), metadata-injected chunking with reverse indexing, and configurable reranking, plus comprehensive eval framework enabling A/B testing across prompts/models/retrieval strategies.
- Architected Commercial Component Engine serving 2,000+ users with 25,000+ annual compliance transactions through RESTful APIs, PostgreSQL backend, and React frontend.
- Redesigned data lake query service, reducing average response time from 8s to <1s (90% improvement) by optimizing SQL queries, implementing caching layers, and refactoring backend architecture.
- Built high-throughput aliasing service handling 70 req/s (14x improvement from 5 req/s) by redesigning with event-driven architecture and asynchronous processing.

PROJECTS

jot — CLI tool for frictionless idea capture; stores and retrieves notes from the terminal without breaking flow.

Companion AI — AI care platform (Next.js, Firebase) serving 200+ NDIS users.

Mars Rover Manipal — Autonomous navigation system. Led AI team to #1 Asia, #8 globally (URC 2019).

More projects at vedwhat.com

TECHNICAL SKILLS

Languages	Python, Java, JavaScript, SQL
Backend	FastAPI, Spring Boot, Django, REST APIs, Microservices
AI/ML	Custom RAG/Agent Frameworks, Azure OpenAI, AWS Bedrock, Vector Search
Databases	PostgreSQL (pgvector), MySQL, Redis
Cloud & DevOps	AWS (EKS, EC2, Bedrock), Azure, Kubernetes, Terraform, Docker, Prometheus
Tools	Git, Jenkins, Grafana, CI/CD

EDUCATION

B.E. Computer Science, Manipal University 2017 - 2021
Minor in Business Studies — CGPA: 9.3/10

LEADERSHIP

- Lead platform architecture for team of 6 engineers, conducting design reviews and mentoring 3 junior engineers as Scrum Master (40% velocity improvement).
- Led AI team for Mars Rover Manipal to #1 Asia, #8 global ranking at URC 2019.