Venkata Mahesh Polumuri

+91 8688716331 | venkat2552004@gmail.com | linkedin.com/in/venkat2552004 | github.com/Venkat2552004
LeetCode: venkat2552004 (1800+ Rating) | CodeChef: venkat2552004

Education

Prasad V. Potluri Siddhartha Institute of Technology — **Vijayawada, A.P, India**
Bachelor of Technology in Computer Science & Engineering | CGPA: 9.12/10 — *2021 – 2025*

Technical Skills

Languages: Python, C++, Java, TypeScript, SQL
Frameworks & Libraries: Model Context Protocol (MCP), FastAPI, Pydantic, Pytest, Spring Boot, Node.js, Express.js, React.js
Databases & Vector Stores: PostgreSQL (pgvector), MySQL, MongoDB, IndexedDB
Cloud & Observability: AWS (Bedrock, S3, Secrets Manager, Lambda, Fargate, EC2, ECS, IAM), OpenTelemetry, Coralogix, Linux/systemd
Core Concepts: Automated LLM Evals, Model Routing, Data Structures & Algorithms, Operating Systems, DBMS, System Design

Experience

Tata Consultancy Services — **Chennai, India**
AI Engineer — *Nov 2025 – Present*

- Developed a 13-stage ETL ingestion pipeline executing automated batch parsing, normalization, and vectorization of **50,000+** enterprise ServiceNow incident logs and SOP documents.
- Configured system prompt caching on AWS Bedrock (Claude Sonnet 5), reducing input token consumption by **≈90%** and reducing projected LLM API costs by **40%** monthly.
- Generated 1024-dimensional AWS Titan embeddings and stored vector records in PostgreSQL using **pgvector with HNSW indexing** for **sub-100ms** similarity searches.
- Integrated Microsoft Presidio for PII masking of **100%** of sensitive records, SHA-256 hashing for duplicate detection, and OpenTelemetry OTLP telemetry for Coralogix observability.

Accenture — **Hyderabad, India**
Software Development Engineer Intern — *Feb 2025 – Jun 2025*

- Engineered backend microservices utilizing Java 17 and Spring Boot, successfully decoupling 3 legacy modules and improving system maintainability.
- Designed and benchmarked RESTful API endpoints handling up to **2,000 requests/minute**, implementing request validation and reducing API response latency by **15%**.
- Integrated Node.js authentication middleware with Java core services using JSON Web Tokens (JWT) and role-based access control (RBAC), securing **10,000+** active user sessions.

Projects

vps-mcp – Production Read-Only Linux Diagnostics MCP Server | Python, MCP, systemd, Pytest — **2026**

- Engineered a zero-dependency Python MCP server exposing **14** system diagnostic tools, `vps-docs://` resources, and incident-triage prompts over stdio JSON-RPC 2.0 on an Ubuntu ARM64 VPS.
- Designed privilege-separated architecture eliminating `sudo` via systemd socket activation (**0660** Unix socket, `NoNewPrivileges=true`), fixed command allowlists, regex secret redaction, and bounded context outputs.
- Implemented an LLM-as-a-Judge eval harness simulating **20+** outage scenarios, achieving **95%** tool-selection accuracy and automated root-cause diagnosis in **<3** tool calls per incident.
- Built a `pytest` CI suite validating secret masking, input sanitization, and end-to-end stdio transport protocol compliance.

Winnow – AI Inbox Triage & Feed Filtering Chrome Extension | TypeScript, Chrome MV3, OpenAI SDK, Vitest — **2026**

- Developed a Manifest V3 extension automating Gmail categorization and LinkedIn feed filtering via a multi-tiered AI pipeline (local heuristics → Roar AI gateway → Sarvam fallback).
- Designed an automated LLM evaluation harness across **70+** test fixtures, achieving **94%** post classification precision and reducing primary LLM API costs by **35%** via intelligent fallback routing.
- Built asynchronous processing pipelines featuring IndexedDB score caching, concurrency/retry queues, and `declarativeNetRequest` header rules to bypass browser origin restrictions.
- Enforced high software reliability with unit testing (Vitest), golden-fixture heuristic evals, and Playwright E2E automation.

Certifications & Achievements

- **Star Team Award (TCS)** – Recognized for taking an enterprise AI Knowledge Base Ingestion Pipeline to production in 60 days.
- **AWS Certified AI Practitioner** – Validated proficiency in AWS AI services, Bedrock models, foundational LLM concepts, and responsible AI implementation.
- **Claude Foundations Certification** – Certified in Anthropic Claude prompt engineering, context optimization, and enterprise LLM integrations.