Shreyas Vispute

Software Engineer | Agentic AI | LLM Agents & RAG

San Francisco Bay Area | vispute.shreyas1@gmail.com | shreyasvispute.dev | shreyas-vispute

SUMMARY

Software engineer with 8+ years building enterprise systems, now shipping agentic AI to production: LLM agents with tool calling, RAG pipelines, MCP servers and evaluation on AWS Bedrock. Built a triage agent used daily by 12 support engineers that cuts production-issue investigation from hours to minutes, and a human-in-the-loop pricing agent for a major US telecom. Python, TypeScript, Java; AWS, Azure.

TECHNICAL SKILLS

- **AI & Agents:** agentic AI, LLM agents, tool calling, MCP, RAG, vector search, prompt engineering, AWS Bedrock, LangChain, LangGraph, evals
- **Evaluation & Safety:** golden test cases, LLM-as-judge, A/B testing, adversarial tests, PII masking, prompt-injection detection, human approval, audit logging, AWS Guardrails
- **Languages:** Python, TypeScript, JavaScript, Java, C#, SQL
- **Backend & Data:** Node.js, Spring Boot, REST and GraphQL APIs, microservices, Kafka, PostgreSQL (Aurora), Redis, DynamoDB, Cosmos DB, MongoDB
- **Cloud & Frontend:** AWS (Lambda, Step Functions, ECS Fargate), Azure (AKS), Docker, Terraform, CI/CD, React, Next.js, Jest

EXPERIENCE

IBM, United States – Full Stack Developer **Jun 2023 – Present**

Client: Major US telecom – *Enterprise Pricing*

- Own the pricing automation end to end, from stakeholder discovery and a proof of concept to an AI pricing agent (Python, LLM) used by sales, with human-in-the-loop gates before any change is applied.
- Automates the legacy pricing client with end to end verification and crash-recovery safeguards, cutting 2–3 days of manual entry to 3–4 hours.
- Built LLM-powered knowledge bases exposed through MCP, so teams can query the commerce and pricing domain in plain language; consolidated docs from 6 repositories into ~30 pages with automated ingestion.
- Modernize a B2C commerce platform, where customers browse the catalog, build carts and place orders: port legacy services to Azure microservices (Java, Spring Boot, AKS, Cosmos DB, Redis, Kafka) with a React/Next.js front end; fix revenue-critical pricing defects and implement a checkout feature.

Client: Global automaker – *Dealer Vehicle-Order Platform (1,500+ dealerships, 50K+ transactions/month)*

- Built and deployed a production-incident triage agent on AWS Bedrock that takes an issue from report to root cause in minutes; used daily by 12 support engineers and business users, about 30–40 investigations a week. Scoped with business and technical users; 7 tools, RAG, ECS Fargate, Terraform and GitHub Actions, with PII masking, prompt-injection detection and audit logs.
- Designed the agent's evaluation framework (golden cases, LLM-as-judge, adversarial tests) and tuned prompts and models through A/B tests on Bedrock, iterating each cycle.
- Led a team of 7 developers from requirements through release (code review, mentoring) and was on the on-call rotation and second-level escalation for production incidents; wrote 20+ domain and architecture documents; introduced automated testing.
- Architected an AI codebase agent on AWS Bedrock across 7 repositories, plus an LLM code-analysis pipeline (OpenSearch Serverless, Terraform) adopted by the entire engineering team.
- Built the event-driven order workflow on AWS Step Functions (13 independently scaling stages) and rebuilt order update, cancel and unassign as state machines with retries and rollbacks, so failed calls stay traceable and recoverable; owned the vehicle-matching integration (5 external API contracts, retry and fallback).
- Resolved production bottlenecks (connection exhaustion, race conditions) with Redis caching and Kafka tuning, cutting API response times from about 5 seconds to 1.5–2 seconds to meet the SLA.
- Rebuilt dealer order validation with parallel checks and plain-English rejection reasons (responses 50% faster); added quota management with concurrency control and rebuilt notifications as staged batches with duplicate checks and classified errors, so dealer emails go out reliably.

Hanlon Financial Lab, Stevens Institute of Technology – Graduate Student Assistant **Oct 2022 – May 2023**

- Led design of a Python/Django, React and MongoDB app for student transcripts, reducing data-entry errors 25%.

NBCUniversal (Fandango) – Software Engineer Intern **Jun 2022 – Aug 2022**

- Built serverless consumer sites (React, Node.js Lambdas, DynamoDB) on AWS, increasing rewards processing 40%; launched an AWS Rekognition image-classification system that removed a third-party dependency and cut costs.

FinIQ Consulting, India – Software Developer **Sep 2017 – Aug 2021**

- Led a team of 4 redesigning the Fixed Income/Equity Connect Suite from C#/.NET and SOAP to .NET Core, REST and Angular, improving response time 75%.
- Built a microservices news app (Angular, .NET/Java) with real-time market data, raising customer satisfaction 20%.
- Built a Django/Google Blockly payoff app (+30% client engagement) and a CMS cutting report generation time 50%.

Hella India Automotive, India – Software Developer **Jan 2017 – Sep 2017**

- Designed a C#/.NET QR-scanner app that eliminated 90% of manual processing; earned a Pat on the Back award.

EDUCATION

Stevens Institute of Technology, Hoboken, NJ – Master of Science, Computer Science (GPA 3.8) **Aug 2021 – May 2023**

University of Pune, India – Bachelor of Engineering, Information Technology (GPA 3.75) **Aug 2012 – May 2016**