

Darshan Vaghasiya

(667) 229-9039 | darshangvaghasiya0@gmail.com | linkedin | github

SUMMARY

Software engineer who builds AI products end to end, Rust services, Python LLM agent runtimes, and TypeScript interfaces. At Dialogica I ship an AI legal assistant for law firms running in production with pilot users at real firms, covering human-in-the-loop approvals, a 150+ tool agent registry, and document, email, and calendar integrations. Earlier work spans distributed microservices, LLM inference pipelines, and cloud MLOps. I use AI coding agents and LLM tooling daily, which shapes how I design tool interfaces and approval flows for the agents I build.

TECHNICAL SKILLS

- **Languages:** Rust, Python, TypeScript, JavaScript, C++, Java, SQL
- **AI & Agents:** MCP servers and tools, LangGraph, LangChain, tool calling, human-in-the-loop design, RAG, prompt engineering
- **LLM Platform:** Anthropic and OpenAI APIs, SSE streaming, tracing and observability
- **Backend & Systems:** Axum, FastAPI, Spring Boot, Node.js, REST APIs, OAuth 2.0, JWT, Kafka, Redis, asyncio
- **Data & Storage:** PostgreSQL, Pinecone, MongoDB, SQLite, BigQuery, Faiss, ETL pipelines, event streaming
- **Front-End & Desktop:** SolidJS, ReactJS, Tauri, HTML, CSS
- **Cloud & MLOps:** AWS, GCP, Docker, Kubernetes, CI/CD, Jenkins, Grafana, model deployment, cloud monitoring
- **Testing & Practices:** System design, data structures, API versioning, load testing, Agile

EXPERIENCE

Dialogica

Santa Monica, CA

AI Engineer

March 2026 – Sep 2026

- Built a human-in-the-loop approval system in Rust and SolidJS gating AI email, document, and calendar actions.
- Designed a tool-provider registry unifying 150+ MCP tools across the Rust backend and Python LangGraph runtime.
- Extended the MCP tool surface as integrations grew, keeping schemas and error shapes consistent across providers.
- Implemented pause/resume and reconnect-safe SSE streaming so approvals and tool results survived dropped connections.
- Integrated iManage DMS and SharePoint over OAuth 2.0, clearing a 71-finding audit against a Microsoft 365 tenant.
- Hardened the LangGraph runtime with argument validation, tool-result truncation, and error recovery for pilot use.
- Traced LangGraph runs in LangSmith, using tool-call and token traces to find failures before pilot users hit them.

Blue Sigma Software Solutions

Baltimore, MD

Software Engineer

Jul 2025 – Feb 2026

- Engineered Java and Spring Boot microservices on AWS, with Kafka carrying events between distributed services.
- Secured REST APIs with OAuth 2.0 and JWT, standardizing authentication and meeting compliance requirements.
- Orchestrated CI/CD on Jenkins, Docker, and Kubernetes, moving releases to rolling deploys with no downtime.
- Instrumented Redis caching and Grafana dashboards, using load tests to catch latency regressions pre-release.
- Authored runbooks and design docs for team-owned services, shortening on-call triage for unfamiliar incidents.

VITG

Ellicott City, MD

Software Engineer

Nov 2024 – Jul 2025

- Shipped AWS Lambda ETL workflows enriched with LLM inference, cutting manual reporting effort by 35%.
- Wired Claude multimodal OCR pipelines into document intake, lowering human review time by 30%.
- Delivered an LLM chatbot joining geospatial data with natural language queries for city-planning teams.
- Tuned LLM inference pipelines for cost per token, throughput, and reliability across multimodal workloads.

University of Maryland, Baltimore County

Baltimore, MD

Research Assistant (Academic)

Jun 2023 – Jul 2025

- Rewrote malware scanning modules in C++ with multithreading and memory tuning, raising detection speed by 20%.
- Improved Python and C++ scanner components for Linux, raising detection accuracy and runtime stability.
- Simulated cyberattacks in Dockerized test environments to stress ML-driven detection under realistic load.

PROJECTS

Personalized Video Recommendation Engine | Python, TensorFlow, BigQuery, Dataflow, Kubernetes, Docker

- Trained ranking models over video and text features in TensorFlow, scored against held-out watch data.
- Served the model on Kubernetes behind a REST API, holding sub-100ms P95 latency under local load tests.
- Indexed embeddings in Faiss for approximate nearest-neighbour retrieval in under 50ms per query.

AI Video Summarization Platform | Python, PyTorch, Transformers, GCP (GKE, Cloud Storage), TensorRT

- Fine-tuned a pretrained summarization model on a 100K-sample dataset, evaluated with ROUGE-L scoring.
- Quantized inference to INT8 with TensorRT, cutting benchmark latency and GPU cost by about 40%.
- Compared generated summaries against reference text to tune decoding and prompt settings.

Ask-Your-Docs — AI Document Q&A | Python, FastAPI, LangChain, Pinecone, Streamlit

- Assembled a document Q&A service on FastAPI and LangChain with Pinecone namespaces per document set.
- Routed queries across OpenAI and Anthropic models with tunable temperature and retrieval depth.
- Scored retrieval against a 30-question eval set, tuning chunk size and top-k against hit rate.

EDUCATION

University of Maryland, Baltimore County

USA

Master of Science in Software Engineering

Jan 2023 – May 2024

Gujarat Technological University

India

Bachelor of Engineering in Computer Engineering

Jul 2018 – May 2022