

Tanishq Kumar

AI Deployed Engineer | Software Architect | Founding Systems Developer

Specialized in Autonomous Agent Architectures, High-Throughput Backends (Python / Rust), & Cloud-Native Infrastructure

✉ connect@tanishqkumar.me | Hyderabad, India

🌐 tanishqkumar.me | 🌐 github.com/tanishq-kumar | 🔗 linkedin.com/in/-tanishqkumar

EXECUTIVE PROFILE

Founding Backend & AI Systems Engineer with 2.2+ years of production experience architecting **enterprise GenAI platforms and agentic AI systems**, distributed microservices, and automated cloud infrastructure. Built the **xflow.solutions agent harness from scratch**: agent frameworks are excellent for prototyping, but owning the requirements and the intent yields a leaner, better-optimized production system. Set up **MCP servers and unified agent tools** that work in-app and over MCP, and engineered custom **AST compilation engines** and column-level data lineage systems, translating business-level data semantics and data contracts into technical models. Proven track record of accelerating pipeline compilation from 15 minutes to **sub-100 ms (~3,000× faster)**, automating Kubernetes clusters via **Terraform IaC** in under 5 minutes, maintaining **SOC 2 Type II compliance**, and sustaining **99.9% production uptime**. Expert in bridging frontier LLM research with high-reliability enterprise backend engineering.

TECHNICAL SKILLS & CORE COMPETENCIES

AI Engineering & Agentic Systems: Custom Agent Harnesses (Built from Scratch), Model Context Protocol (MCP) Servers, Unified Agent Tools (In-App & over MCP), FastMCP, Anthropic Claude API/SDK, OpenAI API/SDK, Function & Tool Calling, Structured Outputs (Pydantic v2), Retrieval-Augmented Generation (RAG), Semantic Vector Search, DSPy, Open-Source LLMs (LLaMA 3, Phi-3), Context Sandboxing, Trace-Level Circuit Breakers, Prompt Caching & Versioning, FaceNet, MTCNN

Observability & Evaluation: Arize Phoenix, Prometheus, Grafana, Custom Instrumentation, OpenInference, OpenTelemetry Spans, TTFT & Latency Profiling, Automated Regression Eval Benches, Dynamic Multi-Tenant Token & Cost Metering, Budget Governance

Backend Systems & Languages: Python, Rust, SQL, Cypher, TypeScript, FastAPI (ASGI), Axum, SQLx, Tokio, Pydantic, SQLAlchemy ORM, Alembic Migrations, PyO3, RESTful APIs, Async I/O, Query Optimization, Caching, Cursor-Based Pagination, Webhook Orchestration, AST Lexers & Recursive Descent Parsers, In-Memory State Machines, Column-Level Lineage, Agentic Coding (Claude Code, OpenAI Codex)

Databases & Data Systems: PostgreSQL (Full-Text Search, GIN Indexing, Connection Pooling), Memgraph, Neo4j (Cypher Query Language), MSSQL, Apache Arrow, Polars (Zero-Copy Streaming), Azure Data Factory, ETL Pipelines, SCD Type 1 & Type 2 Transformations, Data Semantics & Data Contracts, DBever

Cloud, DevOps & Infrastructure: Kubernetes (DigitalOcean DOKS), Terraform (Modular IaC), Helm Charts, Docker (Multi-Stage Build Optimization), VM Administration (DigitalOcean Droplets), Azure DevOps CI/CD (Self-Hosted Runner Pods & VMs), Cloudflare (DNS, Edge Routing, SSL/WAF), Lens (Kubernetes Cluster Management), GCP, AWS (Familiar)

Security, Governance & Quality: SOC 2 Compliance Test Bench, Dependency Vulnerability Scanning (pip-audit, Trivy), Auth0 SSO, OAuth 2.0, Role-Based Access Control (RBAC), Audit Trails, Mypy (Strict Typing), Ruff Linters, Pytest, Linux/Bash, Git

PRODUCTION ENTERPRISE EXPERIENCE

Software Developer (Founding Engineer) | AI Platform & Systems Developer

Xyenta Solution (xflow.solutions), Hyderabad, India

Jun 2025 – Present

- **Built the Agent Harness from Scratch:** Built the agent harness for xflow.solutions from the ground up on the native Anthropic and OpenAI SDKs. Agent frameworks are excellent for prototyping; owning the requirements and the intent let us keep the agent loop, tool execution, and structured outputs fully under our own control and tuned for production.
- **Engineered AI Canvas Execution Engine & Compiler:** Architected the custom formula compiler and Abstract Syntax Tree (AST) execution engine for xflow.solutions, translating business glossary definitions, SQL, and natural language into live DAG nodes, **slashing process build time from 15 minutes to <100 ms (~3,000× speedup)**.
- **Data Semantics, Data Contracts & Column-Level Lineage:** Mapped business-level data semantics and data contracts onto technical schemas, and designed an in-memory lineage state machine traversing the AST directed acyclic graph (DAG), tracking end-to-end data provenance and column mutations across multi-stage ETL operations with zero schema hallucination.
- **MCP Servers & Unified Agent Tools:** Set up Model Context Protocol (MCP) servers (FastMCP) with 10+ custom tools and built a unified tool layer so the same agent tools work in-app and over MCP; engineered an abstract ToolSpec isolating UI side-effects from pure agent context and enforced sandboxed execution stages, **reducing agent failure rates by 40%**.
- **High-Reliability SDK Migration & Determinism:** Migrated core agentic pipelines from DSPy to native OpenAI SDK and Anthropic SDK (Claude 3.5 Sonnet/Haiku), debloating runtime dependencies by 35% and achieving **99% execution accuracy** with deterministic Pydantic v2 structured schemas.

- **Observability: LLM Tracing, Metrics & Dashboards:** Deployed Arize Phoenix, Prometheus, and Grafana via Helm onto Kubernetes; added custom instrumentation across xflow services alongside OpenInference and OpenTelemetry spans, achieving a **p90 tool-execution latency of 50 ms** and profiling Time-To-First-Token (TTFT) across 100% of production LLM requests.
- **Prompt Caching Economics & Token Metering:** Structured static system prompt prefixes and tool schemas to achieve an **84% prompt cache hit rate on OpenAI** (68% across other providers); engineered dynamic multi-tenant token and cost metering per model, user, and team with automated executive cost reports.
- **Backend Services & APIs (Python & Rust):** Built asynchronous FastAPI services (SQLAlchemy, Alembic) with Rust (**Axum, SQLx, Tokio**) on performance-critical paths, and designed APIs to stay fast under load: async I/O end to end, connection pooling, query optimization, caching for hot reads, and cursor-based pagination with streamed responses; cached PostgreSQL full-text search serves at **sub-20 ms retrieval latency**.
- **Automated Cloud Infrastructure (Terraform IaC, Kubernetes & VMs):** Modularized cloud provisioning for DigitalOcean Kubernetes (DOKS) clusters and Droplet VMs using Terraform (IaC), **cutting infrastructure provisioning time from 2 hours to <5 minutes**; administered the VMs, authored declarative Helm charts with Horizontal Pod Autoscaling (HPA), and managed clusters with Lens.
- **Enterprise CI/CD with Self-Hosted K8s Runners:** Designed automated Azure DevOps CI/CD pipelines deploying self-hosted remote agents inside private Kubernetes pods and Droplet VMs, **tripling release velocity** while enabling zero-downtime rolling and canary deployments; optimized multi-stage Docker builds for faster builds and leaner images.
- **Enterprise Security, RBAC & SOC 2 Compliance:** Established an automated SOC 2 compliance bench with execution audit trails, vulnerability scanning (pip-audit, Trivy), Ruff linters, and strict Mypy static analysis, sustaining **85%+ test coverage**; integrated Auth0 SSO/OAuth2 and administered DigitalOcean RBAC via doctl.
- **Technical Leadership & System Uptime:** Configured Cloudflare DNS, edge SSL, and routing to maintain **99.9% production uptime**; mentored 3+ junior engineers on system design, distributed data pipelines, and agentic workflows.

PRIOR ENGINEERING EXPERIENCE

Backend Intern (Product Engineering – xflow.solutions)

Xyenta Solution, Hyderabad, India

Jan 2025 – May 2025

- Prototyped backend microservices for xflow.solutions using FastAPI and PostgreSQL, implementing initial AST formula translation routines to convert business definitions into live process nodes.
- Engineered cached full-text search indexing with sub-20 ms retrieval latency and developed early LLM function-calling tools with Pydantic validation.

Data Analytics Intern (Analytics Team)

Xyenta Solution, Hyderabad, India

Aug 2024 – Dec 2024

- **Enterprise ETL Pipelines:** Designed and maintained automated data pipelines using Azure Data Factory and MSSQL, implementing Slowly Changing Dimensions (**SCD Type 1 & Type 2**) from staging into enterprise fact and dimension tables.
- **Data Lake Transformation Optimization:** Streamlined complex SQL transformation routines, cutting staging data processing latency by 35% and validating data integrity across cross-tenant data models.

Software Engineering Intern (AI & Computer Vision)

National Informatics Centre (NIC), MeitY, Hyderabad, India

Jun 2023 – Jul 2023

- **High-Accuracy Computer Vision Web Service:** Built and deployed a face recognition web service using Python, FaceNet, and MTCNN, processing **10,000+ verification queries** against identity databases at **98.5% recognition accuracy**.
- **Cross-Platform Enterprise Integration:** Architected an inter-process integration bridge connecting Python AI inference routines directly into legacy Java enterprise applications with **<150 ms execution latency**.

KEY SYSTEMS ARCHITECTURE & OPEN-SOURCE PROJECTS

voyager-ogm | *Object-Graph Mapper & ISO GQL Query Compiler (Rust)* github.com/tanishq-kumar/voyager-ogm

- Architected a vendor-neutral Object-Graph Mapper (OGM) and AST query compiler in Rust, providing unified query abstraction across openCypher, SQL:2023 PGQ, and emerging ISO GQL standards.
- Implemented zero-copy streaming query execution powered by Apache Arrow and Polars, compiling complex multi-hop graph queries into optimized relational and graph database execution plans for Neo4j and Memgraph.
- Designed declarative schema macro definitions allowing type-safe entity mapping and traversal definitions with sub-millisecond serialization overhead.

py-layerd & layerd | *Ultra-Fast Sugiyama Layered Layout Core (Rust & Python)* github.com/tanishq-kumar/py-layerd

- Developed a high-performance, low-RAM Sugiyama layered graph layout core in Rust with native Python C-extensions (PyO3) and React Flow / JSON Canvas adapters.

- Engineered offset-aware edge routing and crossing-reduction heuristics, enabling sub-second layout computation for large-scale directed acyclic graphs containing **100,000+ to 200,000+ process nodes**.
- Packaged automated wheel builds via Maturin and UV, enforcing strict memory profiling and achieving zero-leak continuous layout streams.

QAnsiTextViewer | *Qt High-Performance Log Viewer Widget for Python*

PyPI: qansitextviewer | GitHub

- Published open-source package on PyPI (`pip install qansitextviewer`); built a high-throughput GUI log-viewer widget for Python (PySide6/PyQt) featuring streaming ANSI escape code processing, live regex search, bookmarking, and multi-session persistence; benchmarked rendering 100,000+ log lines with zero frame stutter or UI lockup.

DocuSpace | *Document Intelligence & RAG Platform (Smart India Hackathon 2024 Finalist)* github.com/tanishq-kumar/docu-web

- Developed an enterprise document intelligence web platform on GCP (`docu-web`) in TypeScript/React converting non-machine-readable, unstructured files (50+ page PDFs) into strongly validated JSON schemas with live bounding-box OCR highlights.
- Built a Retrieval-Augmented Generation (RAG) pipeline utilizing open-source LLMs (LLaMA 3, Phi-3) with semantic document chunking and vector indexing, delivering **sub-second semantic question answering over 1,000+ chunks**.

Seminar Faculty Planner | *Academic Workflow Platform (MERN Stack, RBAC)* Hyderabad, India | Institutional codebase, not public

- Built and deployed a full-stack academic planning platform with a 2-level hierarchical approval workflow, automated report generation, and role-based permissions, **in production use serving 150+ faculty members across 5 departments**.
- Automated email notifications, schedule conflict detection algorithms, and department-level analytical audit reports.

ARCHITECTURE NOTES: LESSONS FROM PRODUCTION

- Own the Agent Loop:** Wrapper frameworks added dependencies and unpredictability. Wrote an agent harness from scratch over the native Anthropic and OpenAI SDKs rather than on a general-purpose agent framework, replacing an earlier DSPy pipeline with a single `AgentBuilder` factory; constrained decoding with Pydantic v2 schemas reached a **99% function-calling success rate** and removed 35% of runtime dependencies. *Lesson: owning the loop, tool calls, and output schemas is what makes agents predictable.*
- Compile What Has to Be Exactly Right:** LLM-assembled pipelines were slow and could hallucinate schemas. Replaced generative assembly with a custom lexer and recursive descent parser that evaluates typed ASTs as in-memory DAGs, eliminating schema hallucination and cutting build time from **15 minutes to <100 ms**. *Lesson: where correctness is the requirement, a deterministic compiler beats a model.*
- Define a Tool Once, Run It Anywhere:** Tool calls leaked UI side effects into agent context. Built one tool layer serving the in-app agent and external clients (Claude Desktop, Cursor, CLI agents) over MCP; an abstract `ToolSpec` isolates UI side-channels, with idempotent reads and staged transactional diffs for writes, **reducing agent failures by 40%**. *Lesson: a tool is a contract; where it runs is a transport detail.*
- Fast APIs Are Decided Below the Handler:** Endpoints slow down from blocking I/O, chatty queries, repeated reads, and oversized responses. Kept I/O async to the database, pooled connections, optimized queries, cached hot reads, and paged large results with cursors and streaming; cached full-text search answers in **under 20 ms**. *Lesson: get waiting, querying, and payload size right by default and most endpoints never need tuning one by one.*
- Write the Hot Path in Rust:** A few services sat on the critical path. Wrote them in Rust (Axum, SQLx, Tokio) while the rest of the platform stayed in Python on FastAPI, getting **more throughput on less compute**. *Lesson: Python is the right default; Rust earns its place on the hot path.*
- Treat Latency and Cost as Design Inputs:** LLM latency and spend were invisible per tenant. Ran Arize Phoenix, Prometheus, and Grafana via Helm with OpenInference spans and custom instrumentation; stable prompt prefixes reached an **84% prompt-cache hit rate** on OpenAI under per-tenant budget caps and 50 ms p90 tool latency. *Lesson: measure per request and per tenant from the start, and cost becomes something you design for.*

PEER-REVIEWED PUBLICATIONS & RESEARCH

Book Chapter Publication (Wiley):

Kumar, T. (2025). *Optimizing Networking Systems with Machine Learning Approach*. In P. Pavan Kumar, Sandeep Kumar Panda, et al. (Eds.), *Adaptive Artificial Intelligence: Fundamentals, Challenges, and Applications* (pp. 101–133). Wiley.

DOI: 10.1002/9781394389070.ch5

- Researched machine learning applications in telecommunication and network routing architectures, analyzing predictive neural models for packet latency reduction, congestion control, and dynamic bandwidth allocation.

Academic Research in Natural Language Processing (IcfaiTech):

- Collaborated with academic faculty on dependency-parsing structural algorithms for explicit aspect extraction and sentiment analysis across large unstructured consumer review corpora.

PROFESSIONAL CERTIFICATIONS

Neo4j Certified Professional (Neo4j GraphAcademy)

Issued Sep 2026

Credential ID: a89fbe7c-7a18-42c2-ad19-0b83a500f6c8 | Verified Competencies: Graph Data Modeling, Cypher Query Language, Index Optimization, Graph Algorithms, Database Administration

EDUCATION & ACADEMIC DISTINCTIONS

Bachelor of Technology (B.Tech), Data Science and Artificial Intelligence

Aug 2021 – Aug 2025

IcfaiTech, ICFAI Foundation for Higher Education, Hyderabad, India | **CGPA: 9.29 / 10.0** | Merit Scholarship (All 4 Years)

- **Academic Honors:** Awarded Merit Scholarship for Academic Excellence across all semesters (top percentile of class).
- **Core Coursework:** Distributed Systems, Deep Learning, Natural Language Processing, Compiler Design, Database Management Systems, Data Structures & Algorithms, Object-Oriented Analysis & Design, Computer Networks, Cloud Computing.

HONORS & ARCHITECTURAL LEADERSHIP

- **Grand Finalist, Smart India Hackathon (SIH) 2024:** Selected among top national engineering teams for DocuSpace, an automated document intelligence and RAG platform evaluated by Ministry and industry juries.
- **Founding Engineer & Architecture Mentorship:** Spearheaded core technical design decisions at Xyenta Solution, authoring RFCs, establishing CI/CD and testing standards, and onboarding 3+ engineering team members.
- **Active Open-Source Contributor:** Maintainer of graph database tools, Rust compilation utilities, and layout engines utilized by developer communities across GitHub.