

B SHIVAM KUMAR

Chennai, India

+91 7806843443

| shivamkumarb2003@gmail.com

| linkedin.com/in/shivam-kumar-b405b1253

| github.com/DracoShivam

Summary

AI engineer (May 2026) with production experience designing and deploying **agentic LLM systems** (Claude & Gemini APIs, **RAG pipelines**) and **deep learning** models (PyTorch, ViT, EfficientNet) at a quantitative trading firm. Built high-throughput **Python** and **React** backends scaling to **192,000+ concurrent processes**; experienced across the full AI delivery lifecycle – prompt engineering, **LLM evaluation**, structured output, and production integration – within **Agile sprint** cycles. **Azure certified**.

Experience

Futures First Info Services

Jan 2026 – Present

Software Engineer Intern – AI Engineering

Hyderabad, Telangana

- Reduced end-user configuration time by **30%** by engineering a production **agentic LLM assistant** (Claude & Gemini APIs) with a **RAG pipeline** over a vector store, grounding natural-language prompts via **prompt engineering** and **JSON schema** output validation in a latency-sensitive environment.
- Eliminated throughput bottlenecks for **192,000+ concurrent processes** by re-architecting a distributed alert-detection platform **24×** (8,000 → 192,000+) using **RabbitMQ** horizontal scaling and **async Python** (asyncio) multi-threaded workers; validated RAG retrieval quality against ground-truth configurations.
- Achieved **sub-10 ms** end-to-end alert latency by engineering a **C#** market-data gateway multiplexing consumers over a single Lightstreamer connection with async **Protobuf** contracts; diagnosed and resolved message-ordering inconsistencies under high-concurrency load.
- Shipped a **React + FastAPI** real-time dashboard with **WebSocket** streaming and a dual-database (**PostgreSQL + MongoDB**) persistence layer, containerised with **Docker** and delivered end-to-end within a **2-week Agile sprint**.

Projects

Deep Learning Defect Detection | *Python, PyTorch, ViT, EfficientNet, GitHub Actions*

Jul – Aug 2025

- Delivered a **computer vision** inference pipeline (Vision Transformer & EfficientNet) for automated image-based defect detection, achieving **>98% accuracy** on held-out test data; benchmarked inference speed vs. accuracy trade-offs across model architectures to inform production deployment decisions.
- Enforced **LLM/model regression prevention** across every merge by authoring **benchmark-gated CI/CD** scripts in **GitHub Actions**, establishing an end-to-end **MLOps** evaluation loop from training through deployment.

Agentic RAG Code Assistant | *Python, FastAPI, Gemini API, Docker, GitHub Actions*

Aug 2025

- Designed an **agentic RAG pipeline** indexing codebase context into a vector store to ground **Gemini API** prompts with semantically retrieved snippets for context-aware refactoring and auto-generated documentation, enabling iterative prompt-and-deploy cycles across **3 zero-regression deployments**.
- Enforced **structured LLM output** via a **JSON schema** validation layer in **FastAPI** and fully automated **Docker + GitHub Actions CI/CD**, reducing manual release effort to zero.

AI Semantic Recommender | *Python, FastAPI, MongoDB Vector Store, Sentence-Transformers, Docker*

Aug – Sep 2025

- Improved retrieval relevance by embedding a **10,000+ document** corpus into a **MongoDB vector store** using **Sentence-Transformers** for nearest-neighbour semantic search; achieved **95% test coverage** across **40+ PyTest** unit and integration tests with automated **CI/CD** via GitHub Actions.

Spatial Analytics Dashboard | *Python, PyArrow, Next.js 16, TypeScript, Vercel*

Mar – May 2025

- Built a high-performance **data visualisation** layer (**HTML5 Canvas**, **Next.js 16 / TypeScript**) rendering **89,000+ spatial telemetry points at 60 FPS**, demonstrating end-to-end ownership from ETL pipeline through front-end rendering optimisation.
- Cut client payload **99.9%** (15 MB → 2 KB) via a **Python ETL pipeline** (PyArrow, Parquet); deployed **serverless** on Vercel CDN with **O(1)** lookups and zero cold-starts.

Education

SSN College of Engineering

Oct 2022 – May 2026

Bachelor of Engineering in Electrical and Electronics Engineering

Chennai, Tamil Nadu

- Relevant Coursework:** Data Structures & Algorithms, Object-Oriented Programming (OOP), Artificial Intelligence & Machine Learning, Distributed Systems, Internet of Things.

Technical Skills

Languages: Python, JavaScript / TypeScript, Java, C/C++, C#, SQL, HTML / CSS

Frontend: React, Next.js, TypeScript, HTML5 Canvas, WebSockets

AI / ML & GenAI: PyTorch, TensorFlow, Computer Vision (ViT, EfficientNet), Sentence-Transformers, RAG Pipelines, Agentic AI, Prompt Engineering, LLM Evaluation, Structured Output (JSON Schema), Claude API, Gemini API, MLOps