

Sai Narayana B

+91 8248881924

sainarayana2311@gmail.com

GitHub: SaiNarayana-B

LinkedIn: sai23

Portfolio: sai-narayana-portfolio.vercel.app

SUMMARY

GenAI Engineer with 1.5 years of experience building and shipping production-grade AI agent systems at Quantiphi. Architected multi-agent orchestration pipelines serving 2,000+ employees; shipped an open-source codebase analysis tool (patchwork-conventions) to PyPI and MCP Registry. Skilled in multi-agent orchestration, Retrieval Augmented Generation (RAG), NL2SQL, LLM evaluation, and agent SDK integration (Google ADK, Claude, Copilot) on Google Cloud Platform.

EXPERIENCE

Quantiphi

Jul 2025 - Present

Machine Learning Engineer (Applied GenAI)

Bengaluru, India

TEX Agent — Multi-Agent Staffing System (RAG & Parallel Orchestration)

- Architected a **multi-agent staffing system** using **Google ADK** that takes a project ID or role description, dynamically spawns parallel agents for each required role, and returns top candidate recommendations with fit rationale.
- Developed a SoW ingestion pipeline that pulls **2,000+ project documents** from Google Drive, summarizes them via LLM, and indexes into **Vertex AI Vector Search** for the staffing agent's RAG retrieval.
- Reduced pipeline runtime by **95%** (45 min → 5 min) by replacing sequential role processing with **dynamic parallel orchestration**.

NL2SQL Agent — QAssure Platform (Text-to-SQL & MCP)

- Designed a conversational **NL2SQL agent** serving **50+ PMs and business leads**, converting plain-English queries to **BigQuery SQL** via an **MCP server** and returning summarized project health insights.

LLM Caching & Cost Optimization (Prompt Engineering)

- Ran implicit vs. explicit **caching POC** and restructured all agent system prompts, cutting daily inference cost from **\$4,000 to \$2,500** (37% reduction).
- Set up **dynamic LLM routing** across Gemini and Claude model families to pick cheaper/faster models based on task complexity, reducing inference cost across production agents.

LLM Wiki — Internal Knowledge Base (REST APIs & GCS)

- Created **REST APIs** to sync the internal knowledge base (LLM Wiki) with **GCS**, keeping **500+ documents** used by production AI agents up to date.

Keybank SDD Demo (Spec-Driven Development & Microservices)

- Delivered a client demo where AI agents generated specs, built a multi-service backend with frontend, tested, and deployed it end-to-end using **spec-driven development**; presented to KeyBank stakeholders.

Codeaira — Internal AI Coding Assistant (CLI & VS Code Extension)

- Audited and improved the internal AI coding assistant (**CLI + VS Code extension**) by benchmarking against **5 open-source tools**, rewriting **system prompts**, and adding slash-command based skill invocation.

Cypher — AI Developer Analytics

- Engineered a developer analytics tool used by **2,000+ employees** that captures AI interaction logs from **Copilot, Kiro, and Claude Code**, scores prompt quality using **LLM analysis**, and calculates effort-saved metrics per developer.

Quantiphi

Jan 2025 - Jun 2025

Machine Learning Engineer Intern

Bengaluru, India

AI Canvas Agent & Rapid Pitch Agent (Agentic AI & Google ADK)

- Designed the **AI Canvas Agent** that generates solution docs and slide decks for **4 GTM teams**, mapping stakeholders, business impact, and ROI from a single project brief.
- Developed the **Rapid Pitch Agent** that researches a target company's financials, AI/ML initiatives, and competitors, then produces a ready-to-use pitch deck for pre-sales teams.
- Implemented **Google ADK** framework and configured **MCP servers** to connect enterprise data sources with AI agents across **3 internal pilot projects**.

SKILLS

Languages: Python, SQL

AI Systems: Multi-Agent Orchestration, Large Language Models (LLMs), RAG Pipelines, NL2SQL, Prompt Engineering, Prompt Caching, Chain-of-Thought, Few-Shot Prompting, Token Optimization, LLM Evaluation, Agentic AI, MCP Servers, A2A Protocol, Function Calling, Model Routing, Guardrails

Frameworks: Google ADK, LangChain, LangGraph, Claude Agent SDK, Copilot SDK, OpenAI API, Gemini API, Anthropic API, FastAPI, Streamlit

Cloud & Infra: Google Cloud Platform (Vertex AI, BigQuery, GCS, Cloud Run), REST APIs, GitLab

Tools: VS Code Extension Development, FAISS, Vertex AI Vector Search, Embeddings, Git

OPEN SOURCE PROJECT

patchwork-conventions

GitHub · PyPI · MCP Registry · Jun 2026

AST-based codebase scanner that auto-generates CONVENTIONS.md for AI coding agents

- Developed **tree-sitter** AST analysis engine detecting naming conventions, import patterns, error handling, testing frameworks, and API shapes across **5 languages** with confidence scores and real code examples.
- Shipped as CLI tool, **MCP server** (8 tools), and Python API; published to **PyPI** and **MCP Registry**; integrates with Claude Code, Cursor, and GitHub Copilot.

EDUCATION

SRM Institute of Science and Technology

2021 - 2025

B.Tech in Computer Science Engineering

CGPA: 9.17

CERTIFICATIONS

- Google Cloud Professional Machine Learning Engineer
- Google Cloud Associate Cloud Engineer
- Google Cloud Generative AI Leader
- Machine Learning Specialization — Stanford University & DeepLearning.AI