

SARTHAK V. KUMAR

+91-8279436744 — me@sarthakvk.com — linkedin.com/in/sarthakvk — github.com/sarthakvk — India

CAREER HIGHLIGHTS

- Scaled Meta Reality Labs spatial audio data pipeline by 1000x (10K → 10M scenes/run), enabling production-scale ML training data generation through streaming execution and distributed coordination
- Building client-facing AI agents, a RAG engine, and MCP tooling on Google Cloud Gemini Enterprise, including a medical-tutor product
- Built two LLM-powered internal tools at Meta: a team context engine and an infra-debugging assistant that resolved 99% of support queries autonomously
- Improved simulation write-path throughput by 21x with bounded-concurrency object-storage writes and memory-conscious pipeline design
- Reduced distributed ATS sync time by 90% at impress.ai through parallelized backend processing
- Built a transactional email pipeline at impress.ai sustaining 100K+ emails/day on AWS SES/SQS/SNS with automated bounce and spam handling
- Owned end-to-end backend and infrastructure build at Highbreed from scratch — AWS container delivery, CI/CD, relational data modeling, and ELK-based observability/search

EXPERIENCE

Unify Technologies

Mar. 2024 – Present

Software Engineer

AI Platform Engineering — Google Cloud Gemini Enterprise

Apr. 2026 – Present

- **Client-facing AI agents on Google Cloud Gemini Enterprise:** Built production agents and a supporting MCP server for a medical-tutor product, exposing domain tools to the agent runtime for client delivery
- **RAG engine:** Built the retrieval layer grounding the agents in client knowledge bases — content ingestion, indexing, and grounded retrieval with cited sources

Meta Reality Labs

Mar. 2024 – Apr. 2026

Embedded full-time on the Reality Labs audio team, owning spatial audio simulation infrastructure — a pipeline that simulates sound propagation across millions of acoustic scene configurations to generate synthetic training data for audio ML models.

- **Scaled scene simulation pipeline from prototype to production: 10K → 10M scenes/run** — Re-architected workflow execution to stream scene batches, schedule work incrementally, and coordinate multi-worker metadata publishing, removing memory, storage, and single-writer bottlenecks
- **Improved write-path performance by 21x** — Parallelized object-storage writes with bounded concurrency, increasing throughput from 1.16 to 24.44 scenes/sec and reducing 100-scene runtime from 86.54s to 4.09s
- **DAG-based scene configuration engine:** Built dependency graph resolution, cycle detection, topological execution, random sampling, and cartesian-product generation for YAML-driven jobs across millions of scene configurations
- **Cross-team data model and label propagation:** Built structured signal/label handling with stable timestamp precision and end-to-end validation; preserved time-accurate event metadata through signal transformations and was adopted by the Sound Understanding team
- **Reliability and developer productivity:** Added regression and end-to-end test coverage across simulation modes, automated expectation updates, stabilized flaky CI paths, and built run-level metrics and visualization diagnostics for self-serve debugging
- **Team Brain — LLM-powered knowledge aggregation:** Built an internal context engine that ingested Jira, code changes, Google Drive, meeting notes, and plans; synthesized live team context with LLMs and automated weekly summaries, saving 3–4 hrs/week per engineer
- **LLM-powered infra debugging assistant:** Built an agentic assistant that reasoned over failed-run logs, prior issues, and system configuration to recommend fixes; resolved 99% of support queries autonomously

Highbreed Development India

Jul. 2022 – Aug. 2023

Principal Software Engineer

Freelancing marketplace platform connecting businesses with vetted professionals.

- **Owned full backend and infrastructure build from zero** — designed production AWS infrastructure, Docker/ECR-based delivery, autoscaling-backed deployments, and independent service releases
- **Designed core relational data models** for marketplace workflows and unified observability/full-text search into an ELK deployment with an eventually-consistent sync layer
- **Mentored 3 engineers** through structured code reviews and engineering best practices

impress.ai

Jul. 2021 – Jun. 2022

Software Engineer

AI-powered end-to-end recruitment automation platform — candidate screening, assessment, and engagement via conversational AI.

- **Reduced distributed ATS sync time by 90%** — re-engineered the sync pipeline with client-hosted systems by parallelizing every processing stage
- **Sustained 100K+ transactional emails/day** — built the pipeline on AWS SES/SQS/SNS with automated bounce and spam handling, and maintained integrations with Greenhouse and Workday

Highbreed Development India

May 2020 – Jun. 2021

Software Engineer

Freelancing marketplace platform connecting businesses with vetted professionals.

- **Established automated CI/CD and containerized deployments** — built Django backend services and Dockerized the application stack

TECHNICAL SKILLS

Languages	Python (primary), Go, C++, Rust, SQL, JavaScript / TypeScript
Backend & Systems	Distributed systems, API design, DAG-based systems, workflow orchestration, stream processing, concurrent & parallel programming, producer-consumer pipelines, event-driven architectures, memory-constrained design
Data	Relational databases, schema design, query optimization, Elasticsearch, Hive, Apache Parquet, DuckDB, columnar storage, time-series data
AI Engineering	LLM applications, AI agents, MCP servers, Google Cloud Gemini Enterprise, agentic debugging assistants, context engineering, RAG pipelines, prompt-engineering tradeoffs, AI-assisted development workflows; foundational ML
Tools & Technologies	FastAPI, Django, Docker, Kubernetes, Redis, AWS SES/SQS/SNS/S3/EC2, Jenkins, ELK, Apache Airflow, WebSocket, NumPy, Pandas, PyArrow, Git, Linux, GCP, Azure

PROJECTS

pi-memory TypeScript • Persistent memory layer for a coding agent with three retrieval tiers — always-injected, model-selected, and on-demand — with scoped user/project stores and per-session token budgets that bound prompt growth; retrieval or selector failures degrade silently rather than blocking the agent turn	github.com/sarthakvk/pi-memory
pi.nvim Lua, TypeScript • Open-source developer productivity tool integrating the pi coding-agent harness with Neovim, enabling developers to share multi-file editor context with the agent, manage project-scoped sessions, and view review findings as native diagnostics	github.com/sarthakvk/pi.nvim
ndq Zig • Open-source NDJSON filtering command-line tool, with a lexer, parser, and AST for a custom query language; executes queries by traversing the AST, supporting nested fields, array indexes, and boolean expressions	github.com/sarthakvk/ndq
gokey Go • Fault-tolerant distributed key-value store in Go using HashiCorp Raft — implemented the FSM interface (Apply, Snapshot, Restore), cluster membership management, and an HTTP API layer for multi-node deployments with leader election and log replication	github.com/sarthakvk/gokey

EDUCATION

AKTU Bachelor of Technology, Computer Science	Noida, India 2017 – 2021
--	---