Siddharth Kundu

+1 (973) 755 8815 | siddharth.kundu95@gmail.com | linkedin.com/in/siddharthkundu | github.com/sid995

Professional Summary

Senior Fullstack Engineer building AI native products and distributed systems end to end, from product discovery to architecture, shipping, and production hardening. Delivered high scale platforms serving millions of users, engineered sub 50ms P99 microservices, and deployed LLM and RAG pipelines into real user workflows. Strong ownership across NextJS/React, Go/Node/Python services, Postgres/Redis/event systems, and AWS/Kubernetes with a track record of improving latency, reliability, cost, and business outcomes.

Skills

Languages: TypeScript, JavaScript, Python, Go
Frontend: React, NextJS, React Query, Redux, Web Performance, Accessibility
Backend & Systems: NodeJS, FastAPI, Gin, gRPC, REST/GraphQL, WebSockets, Microservices, Event Driven Systems, Job Queues (Kafka, BullMQ)
Data & Infra: PostgreSQL, Redis, Prisma, Query Optimization, Caching, Observability, Tracing, Logging, CI/CD (GitHub Actions, Jenkins)
AI/LLM Production: RAG, Embeddings, Vector Search (Pinecone, FAISS), LangChain, OpenAI, Prompt Orchestration
Cloud: AWS, GCP, Kubernetes, Docker

Work Experience

Senior Software Engineer Sep 2025 – Present
SeekLab New York, US (Remote)

- Architected production RAG system from ground up, designing embeddings pipeline, vector search infrastructure (Pinecone), and retrieval evaluation framework that improved candidate match quality **35%** and reduced time to hire **40%** across enterprise accounts.
- Drove end to end technical strategy for AI candidate matching engine, conducting architecture reviews, evaluating competing approaches (semantic search vs. keyword-based), and building consensus across Engineering, Product, and Data Science on technical direction.
- Engineered high scale PostgreSQL infrastructure handling **10K+ concurrent lookups** with **99.9% uptime**; identified and resolved N+1 query patterns, implemented connection pooling and read replicas, and optimized hot paths to achieve **sub 100ms P95 latency** under load.
- Owned full stack product development (NextJS and Node microservices) across 15+ features, establishing reusable component patterns, API contracts, and observability standards that accelerated team throughput **30%** and reduced production incidents **25%**.
- Established LLM productionization standards including prompt versioning, structured output schemas, function-calling patterns, and evaluation metrics; reduced hallucination rates **40%** and enabled reliable A/B testing across recruiter workflows.
- Built retrieval evaluation framework comparing semantic search approaches (cosine similarity, MMR, hybrid search), establishing metrics and automated testing that improved ranking quality **20%** while maintaining **<200ms** search latency.
- Designed fault-tolerant LLM workflows with retry logic, fallback strategies, structured logging, and real-time monitoring; reduced AI feature failures **60%** and established playbooks for debugging production LLM issues.
- Bridged Product Engineering gap by translating ambiguous business requirements into concrete technical milestones, running technical discovery sessions, and shipping iteratively with tight feedback loops and delivering 8 major features on aggressive timelines without scope creep.

Senior Fullstack Engineer Mar 2021 – Feb 2023
DotPe India

- Led 0→1 architecture and delivery of **AdPro**, an AI-powered analytics platform serving **500K+ SMBs**; owned technical strategy, system design, team coordination, and production deployment—drove **40% improvement in merchant ROI** and **$5M+ in measurable customer value**.
- Defined technical vision for real-time analytics platform, evaluating data processing approaches (stream vs. batch), storage strategies (hot/cold tiering), and compute patterns; built cross team alignment on Go based microservices architecture supporting **20K+ concurrent users**.
- Engineered distributed Go microservices achieving **sub-50ms P99 latency** at scale; profiled hot paths with , optimized memory allocations, implemented efficient caching strategies, and designed horizontal scaling patterns that reduced infrastructure costs **$300K/year**.
- Built production ML pipelines for predictive analytics and anomaly detection using statistical models (time-series forecasting, outlier detection) and custom heuristics; processed **10M+ daily events** to surface automated merchant insights, reducing manual analysis time **70%**.
- Architected event driven infrastructure with Kafka processing **50M+ events/day** at **99.95% reliability**; designed partitioning strategy, implemented dead-letter queues, built monitoring/alerting, and established replay mechanisms for fault recovery.

- Drove cost optimization initiatives through infrastructure right-sizing, query optimization, and caching layer design; analyzed telemetry data to identify waste, migrated hot paths to Redis, and consolidated underutilized services—achieved **40% cost reduction** without performance degradation.
- Designed NextJS/TypeScript onboarding platform with AI driven workflows and progressive disclosure patterns; ran usability testing, iterated on conversion funnels, and shipped features that increased merchant adoption **70%** and reduced setup time **50%**.
- Built Dot Design UI component library with **60+ production-tested components**, comprehensive documentation, and accessibility standards (WCAG AA); enabled **3 new product launches in 6 months** and established design system governance adopted company wide.
- Led cross-functional teams of 4–6 engineers across Web and Mobile platforms, conducting architecture reviews, code reviews, and technical mentorship; improved operational efficiency **70%** and doubled feature delivery velocity through process improvements and technical excellence.
- Mentored 3 engineers to mid level roles through deliberate coaching on distributed systems design, system design thinking, and production best practices; established team knowledge sharing rituals and created internal technical documentation that scaled team effectiveness.

Senior Frontend Engineer Aug 2020 – Feb 2021
Acko Insurance India

- Optimized frontend performance to improve page load times **60%**, increasing conversion **12%** and contributing **$1M** in incremental revenue.
- Built high conversion multi-step purchase flows with real-time validation, driving a **25% add-on attach rate** (3× industry average).
- Reduced funnel abandonment **20%** by improving UX patterns across complex form journeys and error/validation states.
- Integrated tests into CI/CD quality gates, maintaining a **99% pass rate** to keep deployments reliable and fast.

Frontend Engineer Feb 2020 – Jul 2020
Honeybridge India

- Built a real-time video streaming web app (React) supporting **1K+ concurrent users** with modern JavaScript and performant UI patterns.
- Improved streaming UX with resilient state handling, latency-aware rendering, and failure recovery for unstable networks.

Frontend Engineer Jan 2018 – Feb 2020
Vinculum Solutions India

- Developed React/TypeScript dashboards for enterprise e-commerce platforms, delivering complex data tables, workflow-heavy UIs, and real-time sync.
- Improved seller workflow efficiency **20%** by simplifying key operational journeys and reducing time-to-complete core actions.

Education

New Jersey Institute of Technology Sep 2023 – May 2025
Master of Science in Computer Science

- **Relevant Coursework:** Machine Learning, Deep Learning, Artificial Intelligence, Data Structures, Web Systems, Distributed Systems, Systems Engineering

Projects

AI Music Generator | NextJS, Better Auth, Prisma, Python, Modal, OpenAI APIs, AWS S3, PostgreSQL

- Built and shipped a full-stack AI product for real-time music generation with an intuitive UX for complex AI controls (prompt refinement, style selection, regeneration).
- Implemented a Python backend with job queue patterns for long-running inference and robust status/progress handling.

PDF Research Assistant | NextJS, FastAPI, LangChain, OpenAI, PineconeDB, NeonDB, AWS S3

- Developed AI-powered document search with semantic retrieval and conversational chat using a production-style RAG pipeline.
- Built streaming chat UI, embeddings-based vector retrieval, and PostgreSQL metadata storage; optimized for **<1s search latency** and strong error handling.

E-Commerce Microservice Platform | Go, Gin, gRPC, GraphQL, Docker, Kubernetes, PostgreSQL

- Designed and implemented a cloud-native distributed e-commerce platform using microservices architecture with gRPC for low-latency communication.
- Built a GraphQL API gateway to unify service access and prevent over-fetching; containerized with Docker and orchestrated with Kubernetes for resilience.