

Arnav Panigrahi

California, USA | 951-347-2065 | arnav.panigrahi@gmail.com | www.arnavpanigrahi.com | www.linkedin.com/in/arnavpanigrahi

Experience

Software Developer, AllCheer – California, USA Mar 2025 – Present

- Built a HIPAA compliant LangChain agent-driven ROI automation orchestrating stakeholder authorization, compliance validation, and document distribution, enforcing multi-party approval logic and reducing manual processing time by 60%.
- Engineered a containerized FastAPI service that emulated authenticated user sessions against RethinkBH endpoints to retrieve clinical records in real time, bypassing public APIs with 24–48 hour latency for downstream applications.
- Developed a React + FastAPI platform using OpenAI-powered NER to extract structured data from therapist notes, achieving 92% accuracy and enabling backend pipelines for scheduling, analytics, and automation workflows.

Software Developer Intern, AllCheer – California, USA Jun 2024 – Dec 2024

- Designed a high-throughput backend logistics engine in Python on Google Cloud, integrating Google Maps Platform APIs for multi-stop route optimization and large-scale staff scheduling across operational workflows.
- Led rapid backend R&D and API schema design for ROI automation, prototyping core workflow logic with Make.com and translating validated concepts into production-oriented service architecture integrated with internal databases.

Junior Software Developer, Pinnacle Consulting LLC – Bhubaneswar, India Jan 2022 – Aug 2023

- Integrated ArcGIS Online geospatial services with Angular and .NET backend systems for a \$75M utility provider, enabling real-time infrastructure visualization and reducing operational oversight errors by 30%.
- Delivered 3 full-stack ERP systems using Node.js and SQL Server, implementing JWT-based RBAC, backend business logic, and complex query workflows to automate invoicing and improve operational efficiency by 25%.

Technical Skills

Languages: Python, SQL, JavaScript, TypeScript, C#, .NET

AI / ML: LLM Applications, LangChain, LangGraph, RAG Pipelines, Vector Search, Prompt Engineering, NLP, Hugging Face, PyTorch, TensorFlow

Backend: FastAPI, REST APIs, Backend Services, Microservices, Distributed Systems, Node.js (Express), Authentication, RBAC, WebSockets

Cloud / DevOps: Docker, Kubernetes, Google Cloud, AWS, Cloud Run, CI/CD, Containerized Deployment

Data / Infrastructure: PostgreSQL, SQL Server, MongoDB, Redis, Pinecone, Qdrant, Git

Projects

Sabi Agent – Rust Coding-Agent Harness | Rust, Tauri, TypeScript, OpenAI-Compatible APIs 2026

- Built a Rust-based agentic coding harness with OpenAI chat completions, structured event streaming, persistent JSONL sessions, resumable conversations, and tool execution for read, write, edit, bash, search, and workflows.
- Implemented agent safety and extensibility features including interactive approvals for destructive tools, session-only approval bypass mode, skill discovery/invoke, provider configuration, CLI onboarding, and a Tauri desktop shell.

Scalable RAG System for Knowledge Retrieval | LangChain, OpenAI, Qdrant, FastAPI 2026

- Engineered a production-ready RAG backend with config-driven ingestion, semantic chunking, embedding pipelines, and Qdrant vector search, enabling low-latency, context-aware question answering over domain-specific corpora.
- Built AI retrieval workflows with source attribution, retry logic, lazy initialization, and test coverage across mocked unit and real API integration paths, improving system reliability and backend maintainability.

bedtime.ai – Multi-Modal AI Storytelling Platform | Phi-3, EfficientNet, TTS, FastAPI, React 2025

- Built a multi-modal AI assistant that transforms user sketches into narrated stories using fine-tuned EfficientNet classification, Phi-3 LLM generation, and Coqui TTS within a real-time inference pipeline, deployed on Google Cloud
- Developed an end-to-end ML system with dataset preprocessing, model fine-tuning (250 classes), and prompt engineering for Phi-3, orchestrated via FastAPI with optimizations for inference latency, accuracy, and user experience

Education

University of California, Riverside Sep 2023 – Mar 2025

MS in Computer Science

Relevant Coursework: Artificial Intelligence, Data Structures and Algorithms, Computer Architecture, Operating Systems