Shubham Shakya

Shubham Shakya | Kamalbinayak, Bhaktapur - 10, Nepal | shubhamshakya369@gmail.com | +977 9706457575
www.shubhamshakya.com.np | suwubham

Profile

I am a Software Engineer specializing in AI-driven backend systems and data-driven solutions with experience deploying cloud-native architectures, integrating intelligent agents, and designing secure, high-performance infrastructure. My work focuses on applied research in AI and system design, transforming innovative ideas into practical, scalable solutions. I am passionate about building intelligent systems that deliver meaningful real-world impact by improving workflows, automating processes, and enabling smarter informed decisions.

Education

2021 – 2025
Dhulikhel, Nepal

BSc in Computer Science
Kathmandu University

Work Experience

2025/03 – present
Maryland, USA - Remote

AI Engineer
Insyde.ai

- Developed AI agents using the OpenAI Agents SDK, integrating with MCP servers, CRM and loan management systems to automate fintech workflows including loan analysis, refinancing evaluation, lead management, and multi-step workflow orchestration.
- Built AI voice agents using LiveKit and Twilio for inbound and outbound campaigns, with low-latency audio streaming, automated scheduling, tool execution, structured call reporting, and human handoff for lead qualification.
- Implemented a Corrective RAG pipeline over Pinecone with relevance scoring and fallback retrieval, plus evaluation sets and tracing to measure grounding and cut hallucinations in financial Q&A.
- Managed deployment of a multi-tenant SaaS platform on AWS, implementing CI/CD pipelines and leveraging serverless and event-driven services to support scalable production workloads.

2024/07 – 2025/03
Kathmandu, Nepal

Backend Developer Intern
Smart Solar Nepal Corporation

- Architected a serverless IoT ingestion pipeline using AWS IoT Core, Lambda, and Kinesis, capable of storing millions of daily sensor telemetry events across S3 and DynamoDB.
- Built Flask-based backend services exposing REST APIs for multi-tenant device operations, user management, and data access, and implemented a WebSocket streaming layer delivering real-time sensor updates to dashboards with sub-second latency.
- Developed AWS Glue ETL jobs to transform and load 1M+ historical telemetry records from Amazon S3 into Amazon DynamoDB for time-series querying and dashboard access.
- Deployed and maintained backend services on Amazon EC2, and managed PostgreSQL schema design and migrations, achieving 99.9% uptime.

2025/02 – 2025/03
Kathmandu, Nepal

Cloud Apprentice
Adex International

- Received hands-on experience in AWS services and cloud architecture, with focused preparation for the AWS Certified Solutions Architect exam.
- Worked on real-world cloud deployment projects, gaining experience in CI/CD pipelines, networking, security, and infrastructure automation.

2024/04 – 2024/10
Kathmandu, Nepal

AI Fellow
FuseMachines

- Participated in a dynamic 6-month AI upskilling program designed by top global experts, working alongside a diverse cohort of AI fellows in a collaborative environment focused on applied learning and innovation.
- Worked on a hands-on project to build and deploy an integrated real-time system combining speech recognition, speech sentiment analysis, and text summarization, showcasing end-to-end model development, deployment, and optimization.

Skills

Software Engineering & System Design — Python, FastAPI, Flask, cloud-native and distributed backend systems, microservices, serverless & event-driven architectures, multi-tenant systems, JavaScript, TypeScript, Express.js, React

AI / ML Engineering — Agentic workflows, OpenAI Agents SDK, Retrieval-Augmented Generation (RAG), Hybrid search, Re-ranking, Context engineering, LangChain, Pinecone, PyTorch, Hugging Face, Voice AI (LiveKit, ElevenLabs, Twilio)

Data Analysis and Visualization — Pandas, NumPy, Seaborn, Matplotlib, Power BI, Excel

Databases — PostgreSQL, MongoDB, Firebase, Supabase, DynamoDB, AWS RDS, Redis, SQLAlchemy

Cloud & DevOps — Amazon Web Services (AWS) — *EC2*, *ECS*, *RDS*, *SageMaker*, *Lambda*, *Kinesis*, *S3*, *DynamoDB*, *Glue*, *IoT Core*; Microsoft Azure, CI/CD Pipelines, Docker, Nginx, GitHub Actions, Infrastructure Automation.

Publications

Devanagari Script Classification using CBOW Embeddings with Attention-Enhanced BiLSTM
31st International Conference on Computational Linguistics

A Comprehensive Framework for Analyzing Collective Sentiment in Multilingual Social Media Feeds
8th International Conference on Innovative Computing and Communication

Adapting Vision-Language Models for Understanding Hate Speech in Multimodal Content
8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Texts

Projects

Senti-Surfer : A Bilingual Approach to Sentiment Analysis for Social Media
A web extension for sentiment analysis of social media content
Senti-Surfer is a research project that combines the power of web scraping, data analysis, machine learning, sentiment analysis, to create an innovative web app that provides users with a way to gain meaningful insights from social media feeds in multiple languages.
Tech Stack: *React, FastAPI, MongoDB*

CIAO - Code In Art Out
A web app for creating generative art with customizable templates and effects
CIAO is a sleek web app that empowers users to create unique art with a wide selection of customizable templates and image effects. It combines user-friendly design with powerful features, allowing for endless creative possibilities.
Tech Stack: *React, Express, MongoDB, Material UI*

SentryHome
A mobile app that transforms old, unused phones into a comprehensive home surveillance system
SentryHome repurposes old, unused phones into a comprehensive home surveillance system. With advanced AI capabilities including real-time object detection and cloud storage, SentryHome ensures robust monitoring and security for homes and businesses alike.
Tech Stack: *Flutter, Firebase, FastAPI*

Certificates

AWS Certified Machine Learning Engineer - Associate
Amazon Web Services (AWS)

AWS Certified Solutions Architect - Associate
Amazon Web Services (*AWS*)

Microdegree in Artificial Intelligence
Fusemachines

AWS Agentic AI Demonstrated
Amazon Web Services (AWS)

Associate Data Scientist
DataCamp

Associate Data Engineer
DataCamp

Honors and Awards

2024/12/31 **Winner - Jiffy Hackathon 2024**
Awarded first place in the Jiffy Hackathon 2024, organized by Udhyam, for developing Jiffy2.0, a creative solution that enhances food delivery through Gamification and Simulation.

2024/09/26 **First Runner Up - CodeWave Hackathon 2024**
Kathmandu University Computational Mathematics Club
Awarded first runner-up for developing an AI-powered medical translator that simplifies complex medical diagnoses into clear, accessible language, aiming to improve health literacy for health-illiterate populations.

2024/04/14 **Artificial Intelligence Scholarship**
Fusemachines
Selected for the highly competitive Fusemachines AI Fellowship, a fully-funded 6-month program offering hands-on AI training, a Microdegree, and career opportunities.

Volunteering

Nepal Research and Collaboration Center
Technical Team
- Organized NASA Space Apps Challenge - Kathmandu, NRCC Research Training 2025 by managing technical platforms including Canvas and Discord, ensuring smooth participation and engagement for over 100 attendees.
- Assisted in the development of the new NRCC website while maintaining and ensuring the stability and performance of the existing site.
- Supporting research initiatives facilitating learning for over 2500 students and strengthening collaborations with over 20 institutions.

Engineers Without Borders Nepal
Service Learner
- Contributed to the Nepal-Germany Intercultural Collaboration Project, engaging with 20 participants to benchmark low-resource language models for improved linguistic accessibility.
- Designed and created quarterly newsletters to engage the community and showcase ongoing projects. Automated email delivery workflows to improve communication efficiency and reach.

Google Developer Group (GDG) - Kathmandu
Documentation Team
- Developed structured proposals and documentation for flagship events such as DevFest 2025, Build With AI, and IO Extended, streamlining collaboration with sponsors and partners.
- Contributed to community-building efforts by ensuring seamless communication, promoting participation, and supporting knowledge-sharing among local developers.