



ABOUT ME

AI & Full Stack Developer at NexApproach, proficient in Python and AWS, engineered high performance platforms that supported 500+ users. Demonstrated strong problem-solving skills by architecting resilient systems with sub-100ms latency, and successfully integrated AI workflows, enhancing service availability and optimizing operational efficiency.

SKILLS

AI Developer Tools: LLM Integration, Prompt Engineering, AI Agents / Multi-Agent Systems, RAG & Context Retrieval
Backend & APIs: Python, Ruby/Rails (service integration), High-Scale Web Services, Distributed Systems
Code Intelligence: AST/Static Analysis, Indexing Pipelines, Knowledge Graph Concepts, Vector Databases
Reliability: Observability (Logs/Metrics/Traces), Performance Optimization, Scalability, SLOs
Languages & Platforms: Python, TypeScript, Rust (familiar), Git, CI/CD, Remote Collaboration

LANGUAGES

ENGLISH
HINDI
URDU

NIHAL KHAN

WORK EXPERIENCE

NEXAPPROACH

TORONTO, ONTARIO
JAN 2025 - PRESENT

AI & FULL STACK ENGINEER

- Engineered a high-performance Next.js platform with LangChain, supporting 500+ simultaneous users through real-time WebSockets and tuned SSR for fast, consistent response times.
- Architected a resilient AWS setup using Lambda (Node 18.x), API Gateway with custom auth, and DynamoDB, achieving sub-100ms end-to-end latency and around 99.5% service availability.
- Delivered secure, scalable REST APIs with JWT-based authentication and 100 req/min rate limiting, while integrating GPT-4 into cost-optimized, fault-tolerant workflows for production use.

SATHYABAMA INSTITUTE OF SCIENCE & TECHNOLOGY

CHENNAI, TAMIL NADU
APRIL 2020 - MAY 2023

RESEARCH ENGINEER

- Built a CNN-based pipeline in PyTorch for larvae identification, achieving an F1-score of 0.92 by fine-tuning pre-trained ResNet-34 models and integrating LLM-generated explanations for misclassification analysis.
- Automated image preprocessing workflows using OpenCV and Kubernetes, reducing inference latency from 200 ms to 140 ms for real-time edge deployments.
- Stored preprocessed data in AWS S3, achieving 99.8% uptime and reducing storage costs by 18% using parquet compression.

EDUCATION

UNIVERSITY OF WINDSOR

Windsor, Ontario
Sept 2023 - Dec 2024

M. ENG. IN COMPUTER SCIENCE

SATHYABAMA UNIVERSITY

Chennai, Tamil Nadu
June 2019 - May 2023

B. TECH IN INFORMATION TECHNOLOGY

PROJECTS

JUNE - SEPTEMBER 2025

RETRIEVAL-AUGMENTED GENERATION (RAG) AI ASSISTANT

Engineered a RAG-based LLM assistant using LangChain and OpenAI APIs, showcasing potential for financial knowledge retrieval and low-latency decision support. Designed agentic workflows with custom input/output handling, reducing hallucinations by 20%. Integrated lightweight ETL and deployed on Hugging Face Spaces via Docker for rapid prototyping. <https://github.com/nihalkhan2810/NihalLLM>

MARCH - MAY 2025

REACT PORTFOLIO

Built a fully responsive portfolio using React.js, Tailwind CSS, and Framer Motion to showcase ML projects and publications. Integrated GitHub and HuggingFace APIs to auto-sync repositories and deployed the site on Vercel. <https://github.com/nihalkhan2810/react-portfolio>
<https://nihalportfolio-rho.vercel.app/>