

NANDHU SREEDEVI KUMAR

+1 (219) 368-1229 | nandhuskumar246@gmail.com

LinkedIn : www.linkedin.com/in/nandhusk | Personal Portfolio : nandhusk.dev

Medium (Tech Blogs): <https://medium.com/@nandhuskumar246>

AI Engineer with 2.5 years of experience, focused on building practical AI systems using LLMs, agentic workflows, and machine learning. Experienced in taking ML systems from development to deployment, with a focus on performance, reliability, evaluation, and efficient use of compute and model resources. I work mainly on LLM applications and agentic systems, including real-time voice agents, autonomous task execution, and AI evaluation workflows. My experience also includes context engineering, AI safety, GPU computing, distributed training, and production-oriented AI infrastructure.

PROFESSIONAL EDUCATION

MS in Applied AI (2025 - 2026) CGPA : 4.0/4.0

Purdue University Northwest, Indiana

B.Tech. in Computer Science and Engineering (2020 - 2024) CGPA : 3.56/4.0

APJ Abdul Kalam Technological University, India

TECHNICAL SKILLS

- **Programming:** Python (Proficient), C, SQL
- **ML & Deep Learning:** PyTorch, CNNs, deep learning, Transformers, model training, fine-tuning, PEFT (LoRA, QLoRA), hyperparameter tuning, XGBoost, scikit-learn
- **LLM & Generative AI:** GPT, LLaMA, prompt engineering, context engineering, embeddings, RAG, LLM fine-tuning, prompt optimization
- **LLM Evaluation & Observability:** LLM/RAG/agent evaluation, prompt evaluation, tracing, latency and cost monitoring
- **AI Safety & Enterprise AI:** LLM guardrails, responsible AI, human-in-the-loop workflows, auditability, tenant isolation, access control, AWS AI-DLC
- **LLM Optimization:** Token optimization, token cost reduction, semantic caching, model selection/routing, latency optimization, inference efficiency
- **Reinforcement Learning:** Q-Learning, POMDP, RLHF
- **AI Deployment:** FastAPI, Docker, REST APIs, cloud inference, RunPod, Hugging Face, Replicate
- **AI Agents & Agentic Systems:** LangChain, CrewAI, SmolAgents, Strands Agents, Claude Agent SDK, MCP, multi-agent systems, autonomous task execution
- **Retrieval & Search:** Pinecone, FAISS, ChromaDB, BM25, vector search, hybrid search, semantic search, ANN, HNSW, vector indexing, reranking
- **Audio & Speech Processing:** STT, TTS, audio feature extraction
- **Cloud & Infrastructure:** AWS Bedrock, AWS AgentCore, AWS SageMaker, EC2, S3, ECR, EKS, Git, GitHub, Linux, CUDA, MPI
- **Data & Scientific Computing:** NumPy, Pandas, Matplotlib, probability, statistics, linear algebra
- **Computer Vision:** YOLO, OpenCV, CLIP, BLIP, image processing

EXPERIENCE

Junior Software Engineer | *ITSoleX India, Mumbai* (Remote) | Part-time (Mar 2022 – Aug 2024)

Over a period of 2.5 years, contributed remotely as part of the engineering team, collaborating with technical leadership and developers to develop, maintain, and release software applications across different business requirements.

- **Backend & Application Development:** Developed and maintained Python-based web applications using Flask and Django, implementing business logic, REST APIs, data validation, database integration, and application features for inventory and operational workflows.
- **Machine Learning & Computer Vision:** Integrated frameworks like scikit-learn, PyTorch, and OpenCV into application workflows for image and document processing, developing and evaluating models for selected classification and data-extraction tasks and reducing repetitive manual verification.
- **Deployment & Production Support:** Contributed to deploying and maintaining applications on AWS, using Docker and CI/CD pipelines for application builds and releases; supported production troubleshooting, configuration changes, and routine application updates.

Machine Learning Intern | *Unified Mentor, India* (Sep 2024 - Dec 2024)

- Designed and implemented end-to-end machine learning pipelines on datasets containing 15,000+ records, including preprocessing, feature engineering, and model evaluation with measurable accuracy gains.
- Developed and optimized deep learning models in PyTorch, achieving ~18% performance improvement over baseline models.

RESEARCH & PROJECTS

RESEARCH AT PURDUE

Quantum-Enhanced AI for Efficient Medical Image Generation (Sep 2025 - Present)

Mentored by Dr. Tae-Hoon Kim

- Addressed the high computational demands of modern AI image generation by producing realistic synthetic medical images for training and research on low-end hardware.
- Achieved a 5–10× reduction in model size and computational cost compared to standard generative models without degrading the quality of generated images by employing a hybrid quantum-classical architecture.
- Provided evidence that a hybrid quantum-classical approach can reduce operational costs by up to 80–90% without sacrificing generative performance.

Lightweight ML System for Depression Detection Using Speech and Text (Jun 2025 - Sep 2025)

*Mentored by Dr. AmirHossein MajidiRad | (Manuscript in preparation for **arXiv** submission)*

- Addressed the challenge of detecting depression by analyzing both what people say (text) and how they say it (audio) to capture context and emotional tone, while reducing the high computational demands of typical models.
- Designed as a lightweight model capable of running on mobile and low-power devices, enabling accessible mental health screening.
- Outperformed text-only baselines, achieving an F1 score of 0.70 using combined BERT embeddings and MFCC audio features.

APPLIED AI PROJECTS

AI Voice Agent for Automated Task Execution via Phone Calls (May 2026 - Present)

- Developed a production-style AI voice agent using the Google Gemini Live API and Twilio for real-time, low-latency phone conversations and autonomous task execution, with built-in traceability and auditability.
- Designed context-aware conversation workflows that dynamically manage call objectives, conversation history, tool outputs, and transcripts to maintain relevant context and minimize redundant LLM processing.
- Implemented structured task configuration, input/output validation, controlled tool execution, and safety-oriented workflow constraints for reliable autonomous actions.
- Optimized agent workflows around latency, context usage, and model/API efficiency for real-time interactions.
- Link: <https://github.com/Nsk246/ai-call-agent>

Hospital Triage AI | Reinforcement Learning | Multi-Agent System (Apr 2026 - May 2026)

- Built a multi-agent reinforcement learning system that monitors patient vital signs and learns to prioritize monitoring, treatment, escalation, and emergency response across multiple patients.
- Trained a Deep Q-Network for individual patient decisions and a supervisor agent for ward-level resource coordination without hardcoded decision policies.
- Integrated Anthropic Claude API to provide human-readable explanations of model decisions, improving interpretability for clinicians and supporting human-in-the-loop review.
- Designed real-time dashboards for monitoring agent decisions, patient states, and system behavior, supporting AI observability and decision traceability.
- Link : <https://github.com/Nsk246/redesigned-funicular>

ThinkTrace - Enterprise AI Reasoning (Mar 2026 – Apr 2026)

- Built a multi-tenant enterprise AI platform using 4 parallel AI agents to detect logical fallacies, verify claims against 5 live sources, and evaluate reasoning quality.

- Designed context-aware agent workflows with selective retrieval and structured outputs to control unnecessary LLM context and improve response efficiency.
- Implemented content-hash caching and retrieval optimization, returning repeat analyses in ~460ms while eliminating redundant LLM calls and associated token usage.
- Designed a triple-store data layer using PostgreSQL, Pinecone, and Neo4j with organization-isolated namespaces, server-derived tenant identity, and access controls.
- Added production-oriented AI evaluation and observability, including calibrated test cases, tracing, latency monitoring, and quality validation with a 100% evaluation pass rate.
- Implemented JWT + OTP authentication, per-IP and per-organization rate limiting, brute-force protection, and audit-friendly processing to support secure enterprise AI workflows.
- Link : <https://github.com/Nsk246/thinktrace> | Live: thinktrace.nandhusk.dev

AI-Powered Design of Safer, Non-Toxic Drug Molecules [Tox-NIO] (Sep 2025 – Dec 2025)

- Addressed the challenge of toxic drug compounds by developing a molecular editor that actively transforms molecules into safer compounds using Neural Input Optimization (NIO).
- The system reduced the predicted drug toxicity of compounds by ~50% (0.55 → 0.28) on the ZINC-250k dataset through direct optimization of molecular features rather than random search.
- Accelerated the drug design process by over 70%, replacing genetic algorithms with projected gradient descent to streamline the creation of safer therapeutics.
- Link : <https://github.com/Nsk246/Tox-NIO>

Implementation of Scientific GPT from Scratch (Sep 2025 – Dec 2025)

- Implemented a GPT-style language model from scratch, including tokenizer, training loop, and transformer architecture without using pretrained models.
- Scaled training across multiple GPUs using distributed PyTorch and CUDA, improving training throughput and efficiency.
- Demonstrated strong understanding of deep learning fundamentals, sequence modeling, and large-scale training systems.
- Link : <https://github.com/Nsk246/Scientific-GPT-Pre-Training-Only>

AI Image Transformer - HuggingFace SmolAgents Open-Source Contributor (Sep 2025 - Present)

- An intelligent web app that turns ordinary photos into custom art by using a team of AI models working together.
- Developed using SmolAgents with FLUX.1 and GPT systems.
- It automatically analyzes images, generates prompts, and iteratively refines them with user input, producing higher-quality results than standard prompting methods.
- Being prepared for inclusion as an official example in HF SmolAgents repository.
- Link: <https://github.com/Nsk246/AI-Image-Transformer-AI-Agent-GUI->

Intelligent Search System for College Website Content (Feb 2024 - May 2024)

- Built an LLM-powered AI chatbot that provides students with specific answers from university website content, reducing the need to manually search through dense web pages.
- Implemented Retrieval Augmented Generation (RAG) using Python, LangChain, and the Reflex framework to ground responses in retrieved university documents.
- Added basic LLM guardrails and policy-based response validation to restrict answers to retrieved university content, reduce hallucinations, and prevent the system from responding outside its intended scope.
- Link: <https://github.com/Nsk246/Generative-Search-on-College-Website>

CERTIFICATIONS

IBM Data Science Professional Certificate • Certified by *IBM & Coursera* • 2024

AWS Cloud Practitioner • Certified by *Amazon Web Services* • 2023

Validation Number : PN4KBE5K0BB11FWP • AWS ID: AWS05173593

Python for Data Science • Certified by *NPTEL - IIT Madras* • 2023

Generative AI with Large Language Models • Certified by *DeepLearning AI* • 2023

Python Basics for Data Science • *IBM* • 2020

