

Vyoman Jain

MTD (AI) student at Singapore University of Technology and Design

vyoman@outlook.com | [Google Scholar](https://scholar.google.com/citations?user=vyomanjain) | [linkedin.com/in/vyomanjain](https://www.linkedin.com/in/vyomanjain) | github.com/VyomJ

WORK EXPERIENCE

A*STAR | Research Intern | Singapore

Aug 2026 – Present

Working part-time on AI security under [Dr. Dinil Mon Divakaran](#) at the [Institute of Advanced Intelligence and Computing](#)

Apple | SDE Intern | Bengaluru, India

Jan 2026 – Jun 2026

Worked on AI/ML systems to automate the end-to-end triage of software issues across Apple technologies including root-cause analysis, deduplication and fix suggestions, as a part of the Wireless Technology and Ecosystems team

CognitiveLab | AI Research Intern | Bengaluru, India

Nov 2024 – Dec 2025

- Built [Nayana](#), a unified multimodal, multilingual, and multitask AI model family. Developed a custom data generation toolkit for generating synthetic vision datasets in low-resource languages for document understanding tasks
- Produced >2 million pages (2TB+) of data in 22 languages by efficiently scaling the pipeline using Modal, cutting time taken by 90% while also reducing the cost by 50% compared to traditional parallelization with a cluster of VMs
- Built Nayana-OCR, a 500M-parameter fine-tuned vision language model on our synthetic data. It outperformed several contemporary OCR methods, and beat the base model by 50-70% on key metrics across 10 Indic languages
- Developed the M3DR framework to train [NetraEmbed](#) and [ColNetraEmbed](#) – multilingual visual document retrieval models, which outperform baseline models by 150% on cross-lingual retrieval tasks. Introduced the [Nayana-IR benchmark](#), comprising 23 datasets to evaluate cross-lingual and monolingual visual document retrieval capabilities
- CognitiveLab won the [Llama Impact Grant by Meta](#) for Nayana – one of the ten recipients globally in 2025

IIIT Bangalore | Research Intern | Bengaluru, India

May 2025 – Jul 2025

Worked on Language Guided Object Manipulation in 3D Indoor Scenes under [Prof. Viswanath G.](#) Finetuned diffusion models for generating high-fidelity synthetic datasets for training Vision-Language-Action and robotics models

IBM Global Remote Mentoring Project | Open-Source Contributor

Jun 2025 – Jul 2025

Selected by IBM to contribute to [zopen community's](#) open-source project on enabling AI inference on z/OS mainframes

- Successfully [ported the llama.cpp framework](#) to z/OS, achieving a 92% pass rate on the core test suite and enabling private, on-device AI for enterprise mainframes. Built a PoC to demo practical AI use cases and MCP functionality
- Engineered patches to resolve core compilation and architectural challenges like endianness and threading, integrated acceleration libraries to boost token generation speed and prompt processing by 10%

Sarvam AI | Summer Intern | Bengaluru, India

May 2024 – Aug 2024

Interned at Sarvam, a sovereign AI pioneer building the GenAI stack for India. Built the core pipeline for SarvamParse – a tool for parsing unstructured data from PDFs into clean, structured data with extremely high accuracy for use in RAG applications.

- Introduced key new features in the LLM-parsing pipeline for robust parsing error handling and boosting parsing accuracy
- Refactored the codebase reducing its size by over 50%, while implementing unit tests, custom logging and optimizing for speed, cost, and efficiency across 90% of all types of documents
- Deployed the project as a scalable Google Cloud service with an API-endpoint and as a Python package

SELECTED PUBLICATIONS

HyperCluster: Decentralized Large Language Model Inference over Peer-to-Peer Wireless Networks - Poster accepted at [ML4Wireless workshop](#) at **AAAI 2026** (non-archival); Full paper accepted at [DAIS 2026](#)

ReviewArena: A Large-Scale Cross-Conference Dataset and Benchmark for LLM Peer Review - Accepted as **Spotlight** paper at **ICML 2026** [AI for Science workshop](#) (non-archival)

M3DR: Towards Universal Multilingual Multimodal Document Retrieval - Accepted at **AAAI 2026** workshop on [New Frontiers in Information Retrieval workshop](#) (non-archival)

Nayana: A Foundation for Document-Centric Vision-Language Models via Multi-Task, Multimodal, and Multilingual Data Synthesis - [Poster](#) accepted at **CVPR 2025 Workshop** on [Vision Language Models for All](#) (non-archival); [Full paper](#) accepted at **ICCV 2025 Workshop** on [Computer Vision for Developing Countries](#)

Nayana OCR: A Scalable Framework for Document OCR in Low-Resource Languages - Accepted at **NAACL 2025 Workshop** on [Language Models for Underserved Communities](#)

EDUCATION

Master of Science in Technology and Design (Artificial Intelligence), SUTD

Sep 2026 – Present

B. Tech CSE, PES University EC Campus

2022 – 2026

9.51 CGPA | Received highest academic scholarship in all semesters (top 5% students), Immense Contribution Award

GEAR Innovative International School | Grade XII CBSE – 95.4% | School Vice-Captain

2022

Greenwood High International School | Grade X ICSE – 97.2%

2020

SKILLS

- **AI | ML | DL:** PyTorch, sklearn, OpenCV, Transformers, PEFT, llama.cpp, LangChain and LlamaIndex
- **Languages:** Proficient in Python and C; Familiar with Go, C++, JavaScript, R, Java, Shell Scripting and SQL
- **Full Stack Development:** React.js, Next.js, Express.js, Node.js, FastAPI, TailwindCSS, uv, Git
- **Databases:** MongoDB, PostgreSQL, Firebase, MySQL, Supabase, Pinecone, FAISS, Qdrant, LanceDB
- **Cloud:** Docker, Kubernetes, Google Cloud Platform, Microsoft Azure, Amazon Web Services, and Modal for GPUs
- **Big Data** with Hadoop, Kafka, Spark; **Parallel Programming** with OpenMP, CUDA and MPI

PROJECTS

Manimator

A tool to transform research papers and math concepts into stunning visual explanations, powered by AI and the anim engine. Featured in [Hugging Face's Spaces of the Week](#) with 5000+ videos generated in a week

Clinical Trial Protocol Document Generator | Alaiy (AI consultancy startup incubated at NSRCEL, IIM-B)

Built a document generation tool for helping medical researchers in easily drafting clinical protocol trial documents for new medicines using agentic RAG for a Singapore-based startup – [Neuratrial](#)

Vehicular Troubleshooting Assistance Agent | **Bosch University Connect Program**

Built a voice-based multilingual Gen AI chatbot using RAG for providing Vehicular Troubleshooting Assistance (VTA) to field technicians, as part of this 6-month project in a collaboration between [Centre of Cognitive Computing and Computational Intelligence \(C3I\)](#) and Bosch. Deployed an agentic version built using LlamaIndex and FastAPI

HyperCluster | **Capstone Project**

Developed a P2P distributed framework for deploying LLMs on a cluster of edge devices. Implemented a self-organizing P2P network with automatic peer discovery using Iroh. Built a distributed LLM inference layer with model sharding using the HuggingFace transformers library.

EXTRACURRICULARS / ACHIEVEMENTS

GRE Score: 333 / 340 – Valid till 2031

Google Cloud Certified – Associate Cloud Engineer: Certified as Associate Cloud Engineer (ACE) with [Google Cloud Skills Boost](#) covering GCP infrastructure skills in Google Compute Engine, Google Kubernetes Engine, Security, Networking, etc.

Subject Matter Expert | PESU I/O: Taught the 'RAGs to Rich AIs with LLM Agents' course to a class of 27 students as part of a university peer-teaching program. Covered a deep dive into Retrieval Augmented Generation (RAG) architecture, advanced RAG techniques, GraphRAG and creating LLM agents with Langchain, LlamaIndex and Crew.ai

Teaching Assistant | Advanced Foundations of Machine Learning: Assisted [Dr. Arti Arya](#) in teaching course UE23CS342AA1 to 5th semester AI/ML cohort. Responsible for conducting hands-on sessions and assignments, updating teaching materials and mentoring students for their projects.

Quantum Computing

- Participated in IBM Quantum events to learn and build with Qiskit – [IBM Quantum Spring Challenge 2023](#), [IBM Qiskit Global Summer School 2023](#), [IBM Quantum Challenge 2024](#). Certified as a Qiskit Advocate
- Lead organizer for [Qiskit Fall Fest 2024 Extension Event](#) at PES University, selected as one of 100 universities globally to host this IBM Quantum-powered event. Selected again to host an Extension Event for Qiskit Fall Fest 2025, being one of the 150 universities chosen globally from thousands of applicants.

Clubs

- **President, QuantuMania Club PESU ECC:** Led this quantum research community; Conducted workshops, hackathons, lab tours, and guest lectures to foster interest in quantum computing
- **Technical Co-Head, Codechef PESU ECC:** Speaker at workshop to introduce C++ STL to 1st and 2nd year students
- **Tech Domain Member, ACM PESU ECC:** Contributed to 'Hacknight' - a Hacktoberfest event and a workshop on CUDA
- **Quiz Team Member, Quiz Quotient Club PESU ECC:** Organized and set questions for several intra-college quizzes