RAHUL SHARMA

+91 7349485773 — rahulsharmaviswakarma@gmail.com — linkedin/in/rahulsharmavishwakarma
github/rahulsharmavishwakarma — rahulsharmavishwakarma.github.io

Summary — Machine Learning Engineer with a strong foundation in AI, machine learning, and software development. Experienced in developing and fine-tuning AI models for computer vision, NLP. Proven track record in research and practical implementations, seeking full-time opportunities in Software Engineering, AI and machine learning.

Skills

Programming Languages: C/C++, Python, Go, SQL, Rust
Libraries: NumPy, Pandas, TensorFlow, PyTorch, Scikit-Learn, FastAPI, NLTK, LangGraph, MLflow, DSPy, Transformers
Tools: Azure, GCP, AWS, Git/GitHub, Docker, Kubernetes, Linux, VS Code, Hugging Face, Jupyter
Databases: SQLite, MongoDB, Neo4j, Milvus, DuckDB, ClickHouse, MinIO, Falkor
Machine Learning: Deep Learning, Computer Vision, NLP, Transformers, Diffusion Models, Multimodal Models, VLMs

Experience

Zysec AI **Oct 2024 – Present**

SDE-2 AI/ML Engineer *Jan 2026 – Present*

- Designed an OCR-enabled unstructured-data ETL pipeline (hundreds of GBs to TBs) to enable analysis across large corpora.
- Fine-tuned BERT-based models (and SLMs) for specific use cases such as NER extraction and embeddings.
- Fine-tuned vision-language models (VLMs) for OCR (e.g., IBM Granite vision models, SmolVLMs, and Qwen-VL).
- Built optimized retrieval pipelines for LLM answer generation, reducing hallucinations by 95%.
- Built scalable agentic systems integrating multiple tools (including MCP servers) and dynamic workflows for task execution (e.g., automated incident triage, compliance document review, and security analysis).
- Built custom tracing and context-engineering systems with long-term memory to learn user intent and improve agent behavior via continuous learning.

SDE-1 AI/ML Engineer *Jan 2025 – Dec 2025*

- Leveraged a strong hold of software engineering principles to design and architect end-to-end AI pipelines, ensuring seamless integration into large-scale enterprise applications.
- Engineered complex data processing workflows and advanced retrieval pipelines for AI agents, effectively handling and processing massive datasets.
- Led the fine-tuning of diverse models for specialized use cases, including classification for NER, Small Language Models (SLMs), and multimodal Vision-Language Models (VLMs).
- Managed model inferencing and focused on the strategic optimization of model and pipeline performance, implementing robust evaluation pipelines to benchmark and ensure high-quality outcomes.

Associate Generative AI Engineer *Oct 2024 – Dec 2024*

- Developed Retrieval-Augmented Generation (RAG) pipelines to improve information retrieval and response accuracy for cybersecurity applications.
- Utilized LangChain to build agentic applications that support real-time decision-making tailored for cybersecurity needs.
- Optimized large language models (LLMs) to manage memory constraints effectively, enhancing scalability and performance for complex, context-heavy tasks.
- Integrated Neo4j for graph retrieval to improve context handling in LLMs, allowing for enhanced understanding and response in multi-turn dialogues.

HacktivSpace (Community Contributor) **Sept 2024 – Mar 2025**

Machine Learning Engineer (Part-time)

- Working on Retrieval-Augmented Generation (RAG) for context retrieval in EdTech applications, utilizing Milvus and Neo4j to manage and retrieve data effectively.
- Developing conversational memory in LLM applications to maintain context across sessions, improving engagement and continuity in AI interactions.
- Using frameworks such as LangChain and CrewAI to create AI agentic workflows for various tasks, supporting automation and enhanced responsiveness.

HKBK College of Engineering **Aug 2023 – May 2024**

Undergraduate Researcher

- Developed a model to analyze and respond to questions on biomedical images with over 70% accuracy.
- Conducted research on fine-tuning large language models like LLaMA for medical question-answering.
- Designed an R-CNN-based model to extract templates and information from documents for fraud analysis.

Varcons Technologies **Aug 2023 – Sep 2023**

Machine Learning Intern

- Built a predictive sentiment analysis model for stock price prediction, achieving over 80% accuracy.
- Conducted research on various machine learning topics in computer vision and language processing systems.

Publications

Beyond Imagery: AI-Enhanced Diagnostic Assistant for Cancer and Tumor Diagnosis using Radiology Imaging **link**
Authors: Dr. Nandha Gopal S M, Rahul Sharma, Nithin M, Prajwal B R, Prashanth Kalgonda *Mar 2024*
International Journal On Engineering Technology and Sciences (IJETS)

Education

Visvesvaraya Technological University **Bengaluru, India**
Bachelor of Engineering(B.E/B.Tech) in Computer Science and Engineering *Sep 2020 – June 2024*
CGPA: 8.6/10

Mahesh PU College **Bengaluru, India**
12th Standard/PU in PCME *May 2019 – Mar 2020*
Percentage: 78% — Marks: 463/600

MES Public High School **Bengaluru, India**
10th Standard *Apr 2017 – Mar 2018*
Percentage: 88% — Marks: 550/625

Certifications

MongoDB Certified Associate Developer **July 2024**
MongoDB *link*

Oracle Cloud Infrastructure 2024 Generative AI Certified Professional **June 2024**
Oracle *link*

Microsoft Certified: Azure Data Scientist Associate **June 2024**
Microsoft *link*

Deep Learning Specialization **Aug 2023**
DeepLearning.AI *link*

Microsoft Certified: Azure AI Fundamentals **Feb 2023**
Microsoft *link*