Aru Sharma

+91-7452029206 | arusharmazxx000@gmail.com | Portfolio

Education

Panjab University Chandigarh, India
Bachelor of Engineerng in Information Technology *Oct. 2022 – Present*

Little Scholars Kashipur, India
Mathematics and Computer Science; (Central Board of Secondary Education); 95% *May 2019 – Jun 2021*

Experience

AI Engineering Nov 2025 – Present
Deskree Inc. *Toronto, Canada*

- Worked on **Tetrix** and building AI agents for your infrastructure including cloud services like AWS .
- Developed **Tetrix CLI -** a tool to review architecture, and security issues and enforce code quality.

ML Engineering Intern Sep 2025 – Nov 2025
Nannie.ai *London, UK*

- Worked on testing and deploying **SOTA Vision algorithms** for classification, segmentation and pose detection for animals.
- Deployed **OSS text to video generation** models for in-house testing and benchmarking against **Veo3**.

OSS Developer at Google Summer Of Code Jun 2025 – Sep 2025
Mifos Initiative *Seattle, WA*

- Developed a **multi-agent bot** let users know the status of **Jira tickets**, questions related to **Slack discussions**.
- Developed a **full-stack web** application using **FastAPI, NextJs and Firestore** as database and Auth client.

OSS Developer at Summer Of Bitcoin May 2025– Aug 2025
Bitcoin-dev-project *Manhattan , NY*

- Designed and prototyped **AI-assisted coding tools for Bitcoin** using small language models and domain-specific Retrieval-Augmented Generation (RAG).
- Improved **data pipelines** to ingest bitcoin related knowledge from Bitcoin Conference talks, correct and summarize them using LLMs

Software Engineering Intern Sep 2024 – Dec 2024
CNCF WasmEdge *Austin, TX*

- Developed a **RAG** based chatbot for code assistance using **opensource LLMs** with **Wasmedge runtime**.
- Created a pipeline to ingest data from **Github repository**, augmented it using QnA pairs, summary and then embed this into a **Qdrant vector database**.

Publications

Robust Speech Emotion Recognition Across Diverse Datasets: A Comparative Study of Deep Learning and Transformer-Based Approaches for VoIP Applications,16th International Conference on Computing Communication and Networking Technologies, 2025 Accepted for publication.

Projects

Hidden-state-extractor | *vLLM, Pytorch*

- Created a custom plugin in vLLM for hidden state extraction from LLMs for Mechanistic Interpretability.
- Uses Pytorch Forward hooks for extraction and sub-processes for consumption via CUDA IPCs.

Memory-Augmented-Agents | *RLMs, Memory Routers*

- Built a chat agent designed with long-term memory capabilities to make them personalised and adaptive.
- Leverages a Recursive Language Model (RLM) for intelligent retrieval and a Reflection Agent for continuous memory consolidation.

Achievements

- Ranked 15 globally on the **NTIRE Image Dehazing and Denoising challenge** at **CVPR 2024**.