

Raghav Gupta

📍 Chandigarh, India | 🌐 raghav56.tech | ✉ raghavrgupta56@gmail.com | 🔗 linkedin.com/in/raghav56 | 📄 github.com/Raghav-56 | 🐦 x.com/Raghav_rgup

About

AI & DevOps Engineer. Fast learner, curious person. Published working packages on [PyPI](#) and [npm](#).

Active work across agentic systems with RAG, ASR and speech recognition pipelines, computer vision, inference infrastructure, Bash scripting and automation tooling. Experience in fine-tuning LLMs & traditional models; deploying and maintaining web and ML services over cloud platforms.

Education

Bachelor of Engineering in Information Technology

2023 – 2027

University Institute of Engineering and Technology, Panjab University

- CGPA: 8.48/10 (as of Semester 5)
- Reliance Foundation Undergraduate Scholar

Technical Skills

Programming Languages: Proficient: Python, C++, Bash Scripting, Lisp/Racket, SQL; Familiar: TypeScript, C, LaTeX, Rust, Haskell, Scala

Libraries/Frameworks: PyTorch, TensorFlow, MCP, HuggingFace, LangChain, OpenCV, Whisper, FFmpeg, Streamlit, Scikit-learn, NVIDIA NeMo, LLaMA-Factory, WASM

DevOps & Deployment: Git, Docker, Nginx, Caddy, OxMgr, systemd, Oracle Cloud Platform, Google Cloud Platform, GitHub Actions, MLflow, Weights & Biases, uvicorn, FastAPI

Others: Working with CLI; Linux OS; remote & cloud servers (SSH config, key management); native UV (environment and dependency management)

Experience

AI/Research Intern | Design Innovation Centre, UIET

December 2024 – October 2025

- Developed computer vision pipelines using OpenCV, FFmpeg, OpenFace for facial keyframe and Action Unit detection and extraction.
- Contributed to training and evaluation of Voice and LLM models for Multimodal Emotion Recognition.
- Implemented media augmentation pipelines using traditional and GAN-based techniques to enhance dataset diversity and robustness.

Backend Intern | AISOC / PClub UIET

July 2025 – August 2025

- ▶ Engineered the model-hosting layer of [PU-API](#), a unified platform for deployment of AI services and inference endpoints.

AI/MLOps Developer | Lafleurtech

June 2025 – November 2025

- Developed a local ASR transcription pipeline for [SanketAI](#), a meeting manager system, featuring speaker recognition, diarization, noise removal, and post-processing.
- Evaluated and fine-tuned models including Whisper, NVIDIA Parakeet for Indian languages speech recognition.

Projects | Open Source, Independent & Hackathons

2020 – Present

- **Published Tools (PyPI & npm):** [git-surgeon-cli](#) (PyPI): Surgical git history rewriting with message scrub, blob scrub, merge flatten, commit filtering, and automatic backup/undo. [flexoki-mcp-server](#) (npm): MCP server for Flexoki theme generation with Tailwind config generation and framework-specific components (React, Next.js, Vue, HTML).
- **Open Source (GitHub):** Built and experimented with 20+ original public repositories across Python, TypeScript, C++, ML tooling, MCP servers, and automation scripts. Contributed to projects including [openMF Community AI](#) for Mifos ecosystem support.
- **Agentic Systems:** Implemented research-oriented Agentic systems with RAG, MCP, and multiple modal providers. Shipped tools like [RLM-agent](#) (CLI, FastAPI, Web UI) for LLM orchestration; built additional CLI, FastAPI, and Streamlit interfaces with session persistence (memory) and streaming responses.
- **Cloud Deployment, Oracle & GCP:** Deployed multiple Web-services and docker instances on Oracle Cloud, configured with systemd, Nginx, Caddy, and OxMgr, and built an Open-Sora v2 (11B) text-to-video API on Google Cloud Vertex AI with async job orchestration and GCS output pipeline.

Additional Information

Soft Skills: Teamwork & Leadership, Technical Communication & Documentation, Problem-solving, Critical Thinking

Work/Research Interests: Alignment via Interpretability, Inference Infra & DevOps, Systems Programming, Continual Learning, Multi-modality, System Design