

Harikrishna C

harikrishnac005@gmail.com | [+91-6282440799](tel:+91-6282440799) | [LinkedIn](#) | [Github](#) | [Huggingface](#) | [Portfolio](#)

Education

Chinmaya Vishwa Vidyapeeth - BTECH in Computer Science - Kochi, Kerala

Sept 2023 – Aug 2027

- Current CGPA: 9.12/10
- Current Year & Semester: 4th Year, 7th Semester

Experience

DevOps Engineer Intern - KGEN - Bengaluru, Karnataka

May 2026 – Present

- Managed and maintained **CI/CD pipelines** end-to-end across both **KGEN** and **Humyn Labs**, owning build, test and deployment workflows.
- Optimized **cloud infrastructure costs** for Humyn Labs **multimodal dataset annotation** platform.
- Owned **deployment management** across environments, diagnosing and resolving production deployment issues.
- Tools Used: Docker, Kubernetes, AWS, Azure, Jenkins, Terraform, Helm, GitHub Actions, Linux

Intern, Bioinformatics - Karkinos Healthcare Pvt Ltd - Bengaluru, Karnataka

Jun 2025 – Jul 2025

- Built a symptom-to-HPO ID mapping framework that fetches HPO data directly from the **HPO database** and embeds it using **gemini embeddings**.
- Implemented a **hybrid search** pipeline combining **exact/keyword matching** with **FAISS** vector similarity search for searching symptoms.
- Tools Used: Python, Langchain, Langgraph, Pandas, FAISS, Gemini embeddings

Projects

Alo – Cross-Platform Speech-to-Text Dictation App

[Live Demo](#)

- Built a cross-platform **speech-to-text dictation app** that transcribes speech with on-the-fly transliteration for **Manglish, Hinglish**, and **English**.
- Designed an end-to-end transcription pipeline combining **ASR**, transliteration and **LLM-based output correction** to improve mixed-language dictation accuracy.
- Integrated **Composio** with **Cerebras inference** to add in-keyboard agentic task automation, enabling users to trigger actions across third-party apps directly while dictating in Manglish or Hinglish.
- Tools Used: Python, AWS, Docker, Kubernetes, Sarvam API, Cerebras, Composio, Vite, React, TypeScript, React Native, Swift

Memories – Local-First Screen Memory & Semantic Search App

[Github repository](#)

- Built a **privacy-first desktop app** that continuously captures screen snapshots, indexes them **fully on-device**, and answers natural-language questions about everything seen, as a self-hosted alternative to **Microsoft Recall**.
- Added **BYOK** support for any **OpenAI-compatible API** including Cerebras, OpenAI, OpenRouter, and local servers.
- Tools Used: Electron, React, TypeScript, Python, FastAPI, LanceDB, SQLite, sentence-transformers, OpenCLIP, MobileCLIP2, MLX

BFCL Function-Calling Fine-tune – Qwen2.5-Coder-1.5B

[HuggingFace model](#)

- Fine-tuned **Qwen2.5-Coder-1.5B** on the **Salesforce xLAM-60k** function-calling dataset using **Unsloth** LoRA SFT, benchmarked on the **Berkeley Function Calling Leaderboard (BFCL) v4**.
- Achieved **76.19% Non-Live AST** and **58.85% Live AST** (+9.86 over base), outperforming every other **sub-2B model** on the board, with accuracy competitive with 4B–8B models like GPT-4.1-nano and Granite-3.1
- Tools Used: Python, Unsloth, PyTorch, LoRA, QLoRA, MLX, BFCL, Hugging Face Transformers

Edgebind – Cross-Modal Embedding Model

[HuggingFace model](#)

- Developed a **cross-modal embedding model** aligning text, image and audio into a unified **512-dimensional space**.
- Uses **CLIP ViT-B/32** for text and image and **AST** for audio with a trained projection layer aligning audio embeddings into **CLIP's embedding space**.
- Tools Used: PyTorch, OpenCLIP, AST, Hugging Face Transformers, torchaudio

Publications

- **Edgebind: Towards Edge-Deployable Multimodal Embeddings via Audio-Caption Alignment into CLIP Space** - Paper published at **CAISc 2026**.

Skills

- **Programming Languages & Frameworks:** Python, TypeScript, JavaScript, Go, Java, React, Next.js, Svelte, Vue
- **AI / ML:** PyTorch, Tensorflow, MLX, FAISS, Hugging Face, Embedding Models, LLM Fine-tuning, NLP, Diffusers, Langchain, GraphRAG, OpenCV, Sklearn, Numpy, Pandas, Matplotlib, OpenAI API, Vector stores, Pinecone, Weaviate
- **DevOps & Infrastructure:** Docker, Kubernetes, Jenkins, GitHub Actions, Terraform, Helm, CI/CD Pipelines, AWS, Azure, GCP
- **Tools & Platforms:** Git, Ollama, vLLM, Postgres, FFmpeg, Linux, Shell, Supabase, Claude Code, Codex, Opencode

Achievements & Community

- Active open-source contributor with merged contributions to **PyTorch RL**, **Unsloth** and **Hugging Face**; Completed **Hacktoberfest 2025**.
- Member of the MLX Community India, OpenAI codex community kochi, CCCL (Claude code community london), Huggingface community india, Tinkerhub
- Helped host the **Build with AI: Gemma 4 Hands-on Workshop and Deployment Best Practices event** at the Google office, Bengaluru.
- Participated in the **OpenAI Codex Community Hackathon** and **OpenAI Open Model Hackathon**; Selected as a finalist in the **Tech 4 Good Hackathon** hosted by FOSS United.
- Participant of **AQUAVISION AUV** competition of 2024.