Aryasomayajula Ram Bharadwaj

Location: Bengaluru — GitHub — ram.bharadwaj.arya@gmail.com — Mobile: +91-9108832338

PROFESSIONAL SUMMARY

AI Engineer with 6+ years developing production AI systems and leading engineering teams. Independent researcher with published work on model explainability and LLM inference optimization. Experienced architecting scalable solutions for conversational AI and agentic workflows.

TECHNICAL SKILLS

- **Programming:** Python, Scala, Java, R
- **ML/AI:** PyTorch, JAX, Transformers, LangChain, AutoGen, LangGraph
- **DevOps:** Docker, Kubernetes, CI/CD
- **Cloud Platforms:** AWS, GCP
- **Systems:** Redis, Kafka, Microservices, PostgreSQL, MongoDB

EXPERIENCE

Associate Technical Architect - Platform **Nov 2024 - Present**

Quantiphi Analytics, Bengaluru

- Designed AI agent system for automated issue severity classification and escalation management.
- Architected and implemented a conversational AI-agent chatbot for answering sales-related queries at a major telecom company, refactoring legacy systems to a leaner modular implementation.

AI Resident - Lossfunk AI Residency **May 2025**

Lossfunk AI Residency

- Selected among 100 applicants for elite 6-week intensive AI residency program with 10 researchers
- Developed STU-PID, a novel activation steering technique achieving 32% token reduction and improved reasoning accuracy on GSM8K benchmark
- Published research findings: Steering Token Usage with PID Control

ML Engineer - Innovation & Development Labs **June 2019 - Nov 2024**

Musigma Business Solutions, Bengaluru

- Led development of multiple high-impact projects in LLM operationalization, automated trading, and MLOps
- Specialized in backend development, DevOps, and ML engineering across various domains
- Received multiple recognitions including Impact Awards and Star Performer of the Team

HIGHLIGHTED PROJECTS

Wiserank.io - AI-Powered Research Discovery **June 2025**

Creator & Full-Stack Developer

- Built research paper search engine with relevance and 'creativity' ranking algorithms and implemented automatic citation generation feature for uploaded manuscripts

KEY TECHNICAL PROJECTS

Conversational Sales Chatbot using AI Agents **Nov 2024 - Present**

Team Lead, ML Engineer, Backend Developer

- Designed and implemented an AI agent system for automated replies for a telecom client's sales chatbot
- Refactored existing codebase from a proprietary framework to LangGraph and significantly reduced overall codebase size
- Revamped and automated the scheduled RAG data ingestion pipeline, improving retrieval accuracy by 20% and reducing time to first token by 3x

LLM Agent Platform **Dec 2023 - Nov 2024**

Team Lead, ML Engineer, Backend Developer

- Designed and implemented semi-autonomous data analysis platform using AutoGen framework

- Developed automated prompt optimization strategies and integrated RAG support with evaluation framework using ensemble of locally hosted LLMs

High-Velocity Trading Platform **2021 - Dec 2023**

Team Lead, Backend Developer, DevOps

- Refactored backend code and migrated deployment from bare-metal to Kubernetes with automated hyper-parameter search for ARIMA models
- Rewrote legacy trade-signal generation from R to Scala using Akka framework, enabled near real-time metric calculation and portfolio visualization

ML Model Operationalization Platform **2019 - 2021**

Backend Developer

- Developed automatic retraining pipelines for image classification models and designed Java microservices for creating and serving Jupyter notebooks
- Engineered ML model deployment service with canary and blue-green deployment strategies

RESEARCH & PUBLICATIONS

Scaling Laws for LLM-Based Data Compression **July 2025**

Lead Investigator

- Investigated scaling laws for data-compression capabilities of LLMs on text, image, and speech modalities.

Steering Token Usage with PID Control **June 2025**

Lead Investigator

- Novel technique reducing computational overhead in LLMs through activation steering with 32% token reduction on GSM8K.

Understanding Hidden Computations in Transformer Language Models **August 2024**

Lead Investigator

- Investigated internal mechanisms of chain-of-thought reasoning and developed interpretability methods for LLM reasoning.

AWARDS & RECOGNITIONS

AI Alignment Awards - Winner **July 2023**

AI Safety Research Competition

- Selected among 118 global entries for winning research proposal on "goal misgeneralization" in AI systems.

Honorable Mention - Eliciting Latent Knowledge **March 2022**

Alignment Research Center

- Recognized for innovative approach to open research problem in AI safety.

Bronze Medal - Build-on-Redis Hackathon **February 2021**

Redis Labs

- Developed text-to-code search tool using CodeBERT embeddings and Redis Stack for private repository indexing.

Excellence Awards (8x) **2019 - 2023**

Musigma Business Solutions

- 6 SPOT (Star Performer) awards and 2 Impact Awards for technical leadership, innovation, and delivery excellence.

EDUCATION

Bachelor of Technology – Electronics and Communications Engineering **2015–2019**

GMR Institute of Technology, Andhra Pradesh