Karamvir Singh

karamvirh71@gmail.com | Github| LinkedIn | +918283880452

EDUCATION

NIT JALANDHAR
B.Tech.- Electronics&Communication
June 2021 | Jalandhar, Punjab
CGPA 8.86/10

SKILLS

PROGRAMMING
• Python • C++

AI AGENTS
• Multi-Agent Orchestration
• Agent Evaluations
• Tool-Augmented Agents

LLMS
• Retrieval Augmented Generation
• LLM Fine-tuning (Lora, Peft)
• LLM Optimization (MLC, Lorax)
• Responsible AI
• Prompt Engineering
• Vision LLMs

VOICE
• Automatic Speech Recognition
• Text-to-Speech
• Language Models

GENERAL
• Transformers
• Classical ML
• Deep Learning & NLP
• Operating Systems
• OOPs
• ML System Design

INFRASTRUCTURE
• CI/CD Pipelines
• Kubernetes
• FastAPI
• Docker
• Amazon Web Services (AWS)
• Google Cloud Platform (GCP)

PUBLICATIONS

Advancing Educational Insights:
Using Explainable AI to Transform Decision-Making in Education

Published in *International Journal of Research in Applied Science & Engineering Technology*
Featured Research • June 2023 • **Read Paper**
DOI: 10.22214/ijraset.2023.48577

EXPERIENCE

HIGHLEVEL | Staff Engineer - AI
November 2024 - Present | Bangalore, India

- Worked with **Support Team** on developing an **ASK AI Support Bot** for the entire platform to resolve queries of all agencies and sub-accounts. Released **Beta Version** with **91% Relevancy** and **94% Grounding**.
- Working with **Conversations Team** on revamping the agentic flows for **Bots Across Vertical Domain** utilizing **Multi-Agent Routing**, **improving** overall accuracy from **75% to 92%**.
- Working with **Content Team** to build **Highly Copilot** for generating **Content Across Multiple Domains**. Reduced **Multiple Content Generation Platform Integrations** to a **Single Co-pilot**.

UNIFYAPPS | Lead Product Engineer - AI
April 2024 - November 2024 | Gurgaon, India

- Architected a **MultiModal AI-Agentic Framework** using **LLM's**, integrating transformer-based models for cross-modal text, speech, and vision processing. Deployed Agents across **10+ clients** seamlessly integerating their complex workflows and helped in achieving **1M+ in annual cost savings** and **30% reduction in processing times** combined.
- Built voice-enabled **FAQ Chatbot** with custom LLMs, **RAG**, **ASR**, and **Query Refinement**, serving **50K+** monthly interactions.
- Engineered **Business Analytics Tool**: Text-to-SQL models with **domain-specific LLMs** and **semantic parsing**, achieving 92% accuracy on complex queries and supporting multiple database schemas.
- Optimized **distributed training** workflows using **LORA (Low-Rank Adaptation)** and Incorporated **LoRAX**, enabling **adapter loading 15+ models** and serving 10+ fine-tuned models on shared GPUs, reducing serving costs by over 80%.

SPRINKLR | Senior Data Scientist
June 2021 – April 2024 | Gurgaon, India

- Architected a **Conversational AI ecosystem** (Voicebot, Chatbot, RTS, Speech Analytics) integrating **ASR** (Automatic Speech recognition), **TTS** (Text-to-Speech), and **LLM** technologies. Leveraged advanced **transformer architectures** (Wav2vec2, Whisper) with domain-specific optimizations.
- Conversational AI suite helped clients to achieve deflection rate of around **50%**, indicating **50% reduction in call agentic workforce**
- Refined system performance to handle **1500 RPS**(Request-per-Second) via **model quantization**, **batching**, and **distributed inference**, maintaining high accuracy while minimizing computational resources.
- Built an advanced **TTS systems** using **VITS architecture** with average **MOS 4.6**

BLITZJOBS | Data Scientist Intern
June 2020 – August 2020 | Remote, IN

- Designed ML pipeline for candidate classification having database of **1M+** candidates: Did feature engineering, model training (ensemble methods), and deployment. Streamlined for **F1-score 95%**, enhancing recruitment efficiency

XPERTREVIEW SOFTWARE SOLUTIONS | ML Intern
April 2020 – June 2020 | Bangalore (Remote), IN

- Constructed NLP-based chatbot achieving **deflection rate of 40%** using NLTK and ChatterBot; Used intent recognition and dialogue management. Deployed to production with scalability considerations.