Dharssini **Karthikeyan**

AI/ML Engineer

Coimbatore, Tamil Nadu, India

(+91) 7695890560 | dharssinikarthikeyan@gmail.com | Dharssini | dharssini-karthikeyan

"Building production-grade AI systems that bridge research innovation and real-world impact."

Professional Summary

AI/ML Engineer with 2+ years architecting production-grade AI systems across healthcare and enterprise domains. Expertise in multi-agent conversational AI, RAG pipelines, LLM fine-tuning, and scalable microservices. Demonstrated impact: $4,500+ cost savings, 70% manual effort reduction, 45% UX improvements. Seeking Senior AI Engineer/Research Scientist roles to drive innovation in AI systems and research.

Key Achievements

2024	**$4,500+ Operational Savings,** LLM-powered document retrieval system with QLoRA fine-tuning reducing manual effort by 70%	*RNDsoftech*
2025	**80% Latency Reduction,** Multi-provider geocoding with intelligent caching (15s → 2s) achieving 95%+ accuracy	*CyberTranscend*
2025	**45% UX Enhancement,** PDF source highlighting with ML-based out-of-scope blocking for enterprise chatbot	*Improva*
2025	**32% Accuracy Improvement,** Excel-to-database automation using hybrid LLM+ML approach overcoming data irregularities	*Improva*

Education

Coimbatore Institute of Technology — *Coimbatore, Tamil Nadu*

M.Sc., Decision and Computing Science (Five-Year Integrated) | CGPA: 9.59/10.0 — *Sep. 2020 - May 2025*

Technical Skills

AI/ML	LLMs (GPT, LLaMA, QLoRA Fine-tuning), RAG Systems, LangChain, LangFlow, Multi-Agent Systems, PyTorch, TensorFlow, Transformers, NLP, Prompt Engineering
ML Engineering	MLflow, DVC, Triton Inference Server, Model Deployment, Experiment Tracking
Vector & Search	Qdrant, FAISS, Apache Solr, Hybrid Search (RRF Fusion), 768-dim Embeddings
Languages	Python (Expert), C++, SQL, TypeScript (Beginner), Shell Scripting
Backend/APIs	FastAPI, Node.js/Express (Beginner), REST APIs, WebSocket, Microservices, Docker
Databases	PostgreSQL, MongoDB, SQLite, Redis, Vector Databases
Cloud/DevOps	Azure (ACI, App Service, Kubernetes), GCP, Docker Compose, Git, CI/CD
Data Science	Pandas, NumPy, Scikit-learn, Time-Series Forecasting, Power BI, Excel (Advanced)
Computer Vision	OCR (OCDRNet, NVIDIA TAO), Facial Recognition, Object Detection
Frontend	React (Beginner), Streamlit, Web Components

Professional Experience

Improva — *Remote / Coimbatore*

AI/ML Engineer — *Apr. 2025 - Present*

- **Multi-Agent AI System:** Built LangFlow/LangChain orchestration routing 7 specialized agents (support, sales, information) with confidence-based logic (70%/50%/30% thresholds) for healthcare applications; engineered hybrid vector search with 3 fusion strategies (RRF, weighted, cascade) across 768-dim embeddings
- **RAG Pipeline:** Developed sophisticated context management processing with conversation histories, knowledge base contexts, and FAQ integration
- **Analytics & Dashboard:** Built React dashboard with JWT auth, RBAC, 17+ real-time visualizations (Recharts), event-driven tracking, and Excel/CSV export for executive reporting
- **Enterprise AI Platform:** Architected production-ready AI assistant with React/TypeScript embeddable chatbot, Node.js/Express backend, 5 containerized microservices, Azure deployment (ACI/App Service/Kubernetes), and Qdrant vector database for knowledge base management
- **Additional Projects:** Location-aware chatbot with web search integration; Excel-to-database automation using LLM+ML (32% accuracy improvement); PDF highlighting with out-of-scope blocking (45% UX enhancement)

CyberTranscend *Remote, Sweden*

System and Research (Freelance) *Jan. 2025 - Present*

- **Mental Health AI Platform:** Engineered full-stack platform with WebSocket voice + REST chat interfaces, microservices (FastAPI gateway, Faster Whisper ASR, TTS), SQLite DB with 34 psychological dimensions (IDG, KEDS), FAISS-powered RAG, and 5-tier crisis detection (suicide, self-harm, violence) across 40+ keywords with emergency protocols
- **Route Optimization:** Developed Google OR-Tools VRP solver with adaptive strategies (PATH_CHEAPEST_ARC/<50, PARALLEL_CHEAPEST_INSERTION/large-scale), OSRM API integration, capacity/time-window constraints, multi-provider geocoding (Nominatim→Photon→MapBox) with caching reducing response time 80% (15s→2s, 95%+ accuracy)
- **Database Architecture for Food App:** Designed hybrid system (PostgreSQL, MongoDB, Redis, Apache Solr) with full-text search and real-time indexing for personalized health recommendations

RNDsoftech *On-site, Coimbatore*

AI Engineer Intern *Dec. 2023 - Mar. 2024*

- **OCR Pipeline:** Automated document OCR using OCDRNet (NVIDIA TAO) on Triton Inference Server with optimized preprocessing for higher accuracy
- **Document Intelligence:** Enhanced classification with LayoutLM (512-token support), fine-tuned with RAG-based enrichment, and deployed using Ngrok/IIS/NSSM/FastAPI
- **Retrieval System:** Built LLM-powered QLoRA-tuned retrieval engine with domain-adaptive prompts, reducing manual document analysis by 70%
- **Business Impact:** Delivered $4,500+ estimated savings via reduced manual handling and faster prototype turnaround

Samsung PRISM *Remote*

R&D Intern *Aug. 2023 - Apr. 2024*

- **Data Tools:** Built Python/Tkinter GUI for KPI visualization from log files, improving UX and reducing manual analysis time by 50%
- **Reporting System:** Implemented automated extraction + Excel reporting pipelines, increasing data-processing efficiency by 40%

Buckman *On-site, Chennai*

Data Science Intern *Jun. 2023 - Oct. 2023*

- **Time-Series Forecasting:** Predicted product sales volume using 9 forecasting models to optimize inventory and reduce operational costs
- **Model Accuracy:** Improved MAPE by 20–30% with CLV-based segmentation and achieved 10% accuracy gain via Matrix Profiling

Amazon *Remote*

ML Summer School Mentee *Jul. 2024 - Aug. 2024*

- **ML Foundations:** Trained by industry experts in data analysis, model development, and practical ML workflows across hands-on projects

Research Engineering & Technical Innovation

Algorithm Design: Confidence-threshold routing algorithm (3-tier: 70%/50%/30%) for multi-agent orchestration · Metadata-aware cascade retrieval pipeline (6-tier fallback) · Hybrid fusion search strategies (RRF, weighted, cascade)
Safety & Alignment: Crisis detection system (5-tier risk assessment, 40+ keyword lexicon) · Rule-based + semantic safety filters · Grounded generation with RAG evidence checking
System Architecture: Designed 7-microservice mental health AI platform · Async worker architecture for parallel RAG calls · Containerized deployment with health monitoring
Optimization: 80% latency reduction via intelligent caching · Context management for 8K+ character conversation histories · Dynamic service time calculation for route optimization
Evaluation Systems: Custom RAG evaluation suite (recall@k, faithfulness, context compression) · Built load-testing framework for multi-agent pipelines · Hallucination detection benchmarking

Leadership & Certifications

2020-21 **GeeksforGeeks,** Campus Mandri & Treasurer at GFGCIT (2020-2021). Organized workshops and events.
2021 **DeepVision.AI,** Completed AI for Assistive Tech course. Built Sign Language Detector on Heroku using Streamlit.
2021 **Google Cloud,** Achieved 1st milestone (8 quests, 4 badges) during Google Cloud Facilitator Program.
May-Dec21 **Mentorship - Tekie,** Mentored text-based coding through storytelling for 50+ kids aged 10 and above.

Research Interests & Technical Vision

Autonomous multi-agent systems · Retrieval-augmented reasoning · Long-term memory architectures for LLMs · Safety alignment for therapeutic AI · Embedding optimization & efficient vector search · Production inference systems · Explainable AI for healthcare · Hallucination mitigation strategies · Context-aware conversational AI