KEERTHANA B V

AI Agent & LLM Applications Developer
+91-9901724479 | keerthana.b.v.codes@gmail.com | Bengaluru, Karnataka, India
LinkedIn | GitHub | Portfolio | Research

PROFESSIONAL SUMMARY

AI Agent & LLM Applications Developer with an MCA, specializing in fine-tuning and evaluating NLP models and building production RAG/agent systems. Fine-tuned a Legal-BERT model using LoRA, achieving 84.9% F1 on the CUAD benchmark (Best Paper Distinction, NCRIE 2025). Experienced across the full LLM application stack from data pipelines and model evaluation to RAG retrieval, voice AI, and secure full-stack web development.

TECHNICAL SKILLS

- **Languages:** Python, JavaScript (ES6+), SQL
- **AI & Machine Learning:** Generative AI, Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), Embeddings, LangChain, OpenAI API, BERT, NLP, FAISS, Prompt Engineering, Hallucination Mitigation
- **Frontend & Backend:** React.js, Next.js, REST APIs, PostgreSQL, MongoDB, HTML5, CSS3
- **DevOps & Tools:** Git, GitHub, Linux (Nginx), Vercel, Netlify
- **Methodologies:** Agile, Scrum, SDLC, Unit Testing, Code Review, Technical Documentation

WORK EXPERIENCE

Full Stack Developer | **ASPL Tech Solutions Pvt. Ltd.** | Bengaluru, India *October 2025 – March 2026*

- **Developed and maintained** responsive web applications using React.js and Node.js, successfully delivering client-facing features for the healthcare and retail sectors on schedule.
- **Directly collaborated with 5+ clients** to gather technical requirements, provide regular project updates, and ensure the final software deliverables aligned with their business needs.
- **Contributed to a custom HRMS platform** by building an automated employee onboarding module, which helped reduce administrative data-entry time.
- **Implemented secure authentication** by integrating Role-Based Access Control (RBAC) and JSON Web Tokens (JWT) for secure user logins and data protection.
- **Assisted in application deployment** by pushing code through CI/CD pipelines and helping host production sites on cloud environments.

MERN Stack Developer Intern | **Dyashin Technosoft Pvt. Ltd.** | Bengaluru, India *November 2024 – January 2025*

- Developed a full-stack e-commerce platform using MongoDB, Express.js, React.js, and Node.js (MERN) featuring secure JWT-based authentication and role management.
- Enhanced frontend performance using React Hooks and state management best practices, improving page load speed by 25% and delivering a seamless cross-device user experience.
- Designed and documented 15+ RESTful API endpoints with senior engineers for inventory tracking and order fulfillment, following OpenAPI specifications.
- Resolved 15+ critical bugs during User Acceptance Testing (UAT), reducing the post-launch defect rate by 40% and significantly improving platform reliability.

KEY PROJECTS

AI-Powered Conversational Voice Bot & Google Sheets Integration

Tech Stack: Vapi AI, Twilio API, Make.com, OpenAI GPT-4o-Mini, Google Sheets, Webhooks, JSON Schema.

- **Generative AI Voice Agent:** Deployed a multilingual conversational voice agent (English/Tanglish) using Vapi AI and OpenAI GPT-4o-Mini to automate customer order intake and L1 FAQs.
- **Google Sheets Integration:** Built a serverless pipeline in Make.com using custom webhooks to automatically extract structured call data (name, order details, phone number) and log it directly as new rows in Google Sheets.

- **Automated SMS Workflows:** Integrated Twilio API to instantly send UPI payment links via SMS upon call ending, adding a custom error-handling route to prevent API delivery failures from stopping the order pipeline.

AI Legal Document Intelligence Agent **(scholar.google) (GitHub)**

Stack: Legal-BERT, PyTorch, Hugging Face, OpenCV, spaCy, Python, React.js, MongoDB

- **Fine-tuned** Legal-BERT using **LoRA (Low-Rank Adaptation)** for parameter-efficient training, freezing base weights and training only rank-decomposition matrices achieving 84.9% F1-score and 84.7% accuracy on the 510-contract CUAD benchmark in ~3 hours on a single T4 GPU.
- Engineered a data pipeline using **Pandas** and **Hugging Face Datasets** to parse CUAD's character-span annotations across long-form contracts, applying sliding-window chunking to preserve context across sentence-level labels.
- Tracked **precision**, **recall**, and **confusion** matrix analysis (beyond raw accuracy) to address class imbalance in clause detection, surfacing this analysis in a live interactive dashboard on the React frontend.

E-Commerce AI Customer Support Agent

Stack: ***Python, FastAPI, LangChain, Llama-3.1 (Groq API), Hugging Face, HTML/CSS/JS, FAISS***

- Built a production-grade AI Customer Support Agent using a client-server architecture with a Python FastAPI backend and a responsive HTML/CSS/JS frontend; implemented a RAG pipeline utilizing FAISS vector search and recursive chunking (500-char chunks, 50-char overlap) to retrieve relevant policy rules.
- Validated retrieval accuracy and system safety through a structured manual evaluation suite covering in-scope queries, boundary edge cases, and out-of-scope questions, verifying that semantic search outputs aligned with source documents and system prompt guardrails successfully blocked off-topic inputs.
- Developed simulated enterprise features including CRM context injection (names, order IDs) into the LLM prompt for personalized responses, and an automated ticketing system that intercepts complex queries, generates ticket IDs, and logs chat history to a CSV database for human escalation.

EDUCATION

Master of Computer Applications (MCA) | *August 2025*

RV Institute of Technology and Management, Bengaluru | **CGPA: 8.2 / 10.**

Bachelor of Computer Applications (BCA) | *September 2023*

Community Institute of Commerce and Management, Bengaluru | **CGPA: 8.4 / 10.0**