Priya Malemath

Software Engineer
(+91) 9019529770 | mpriya1043@gmail.com | priya-cse | priya-malemath

Generative AI & **Software Engineer** who has engineered and deployed AI solutions across NLP and data engineering. Fine-tuned LLMs to improve prediction accuracy by **25%**, and built Retrieval-Augmented Generation (RAG) pipelines with vector database integration, reducing semantic search latency to under **500 ms.**

Education

KLE Technological University, Hubballi **Jun 2024**
B.E. Computer Science Engineering, GPA - 8.56
Main courses: Machine Learning, Deep Learning (Neural Networks), Exploratory Big Data Analytics

Skills

Programming & Frameworks -
Proficient: Python, C++, SQL, TensorFlow, PyTorch, Keras, Vector Store DB (Pinecone), NLP
Hands-on Experience: Node.js, React.js, Vercel AI, JavaScript, TypeScript, HTML, CSS, Playwright, MongoDB
Currently Learning: Azure OpenAI
Generative AI, Tools & Data Platforms-
Proficient (Gen AI): Hugging Face Transformers, LangChain, RAG, Prompt Engineering, Agentic Workflows, Fine-tuning LLMs, LLM API Integration (OpenAI, DeepSeek, etc.), Azure CI/CD (Jenkins)
Proficient (Data Platforms): PostgreSQL, Git/GitHub, Jupyter, VS Code, Cursor

Experience

Intern
MediCodio, Bengaluru **Jan 2024 – Jun 2024**
- Prototyped ETL pipelines using **SciSpacy** and **BioBERT** to clean and normalize 1K+ clinical records, achieving a **20%** reduction in preprocessing time.
- Benchmarked LangChain vs. Pinecone embeddings (text-embedding-3-small vs. text-embedding-ada-002), selecting the pipeline that boosted semantic precision by **25%**.

Associate Software Engineer (*Promoted from Software Engineer Trainee)*
MediCodio, Bengaluru **Jun 2025 – Aug 2025 | Jul 2024 – May 2025**
- Fine-tuned existing transformer APIs (e.g., OpenAI) on proprietary medical datasets, boosting the prediction of medical-code accuracy by **25%**.
- Engineered advanced prompt-engineering strategies and integrated Retrieval-Augmented Generation (RAG) workflows with Pinecone vector store, delivering **<500 ms** semantic-search response times.
- Automated full-stack UI validation using Playwright, slashing regression testing cycles from **48hr to 4hr**.
- **Co-led** architecture redesign with CEO, reducing codebase complexity by **30%** and accelerating new-feature delivery.
- Worked closely with clients to ensure timely delivery of requirements, consistently meeting **24-hour** turnaround.

Projects

Pinecone RAG vs OpenAI Chatbot
Typescript, Next.js, Pinecone, OpenAI API, Vercel AI SDK **Jun 2025**
- Developed a production-grade RAG chatbot, reducing hallucination rate by **35%** and ensuring **99.9%** uptime through real-time monitoring.
- Automated ingestion and embedding of **100+** web pages into Pinecone via a custom crawler, achieving **sub-150 ms** vector lookup latency for high-speed semantic retrieval.

Doodle-Recognition
Deep Neural Networks, Machine Learning, OpenCV **Jun 2024**
- Trained a CNN on **30,000** OpenCV-processed doodles (**10 classes×3,000 each**), achieving **95%** test accuracy.
- Built **dual input modes**—camera-scan & in-air capture—using OpenCV video streams and contour detection.
- Exported feature embeddings for scalable sketch retrieval and downstream ML integration.

Enhancing PSO Clustering with Autoencoders
Machine Learning, Deep Learning **Aug 2023**
- Automated k-selection via the Elbow method on **five stock-market metrics** (High, Low, Close, Open, Volume).
- Applied an autoencoder to reduce dimensions—DB-Index **0.98→0.50**, Silhouette **0.05→0.66.**
- Accelerated PSO convergence by **~40%**, outperforming baseline across all feature sets.