RAVIVARMA PATTURI

Bengaluru, India | patturiravivarma@gmail.com | +91 8919411528
linkedin.com/in/ravivarma-patturi | github.com/ravivarmapatturi | neuralmasteryai.com

AI/ML Engineer · Generative AI · Agentic & Multi-Agent Systems · LLMs · RAG · NLP · 4 Years

SUMMARY

AI/ML Engineer with 4 years of experience designing and shipping production-grade Generative AI, Agentic AI and NLP systems — LLM-driven multi-agent platforms, knowledge-graph reasoning, RAG pipelines and deep learning models — for leading U.S. healthcare payers and providers and for product-focused AI companies. Took a benefits-digitization PoC from zero to a production project in 14 days, and built a multi-agent claims-adjudication system that cut adjudication cost by 80%.

TECHNICAL SKILLS

GenAI & Agents: LLMs, Prompt Engineering, RAG (hybrid retrieval, cross-encoder reranking, query rewriting), LangChain, LangGraph, LlamaIndex, Multi-Agent Systems, ReAct, Plan-and-Execute, Recursive Language Models (RLM), Knowledge Graphs, Fine-tuning (LoRA)

ML / DL: PyTorch, TensorFlow, Hugging Face Transformers, ModernBERT, RNN/LSTM/GRU, CNNs, Autoencoders, XGBoost, scikit-learn, NLTK, spaCy, Quantization & Inference Optimization

Backend & MLOps: Python, FastAPI, Temporal, Docker, Kubernetes (AWS EKS), AWS, Azure AI Search, ChromaDB, Git/GitHub, Linux, Streamlit

Languages & Data: Python, C, C++, MySQL, Pandas, NumPy, Matplotlib

WORK EXPERIENCE

AI Engineer — Autonomize AI **Feb 2026 – Present**

Genesis AI Studio — platform for building & deploying AI agents

- Designed and implemented the Recursive Language Model (RLM) architecture, enabling hierarchical reasoning through Master–Sub LLM collaboration.
- Researched and integrated the RLM strategy into the agent platform alongside ReAct and Plan-and-Execute workflows.
- Built the orchestration workflow on Temporal for scalable, fault-tolerant execution of production AI agents.

Healthcare AI — Benefits Digitization

- Delivered a proof of concept for a leading U.S. healthcare organization within a 14-day timeline; its success converted it into a full-scale production project.
- Leading development of the production solution — hardening and scaling the extraction pipeline from PoC to a production-grade system.
- Built LLM + Knowledge Graph pipelines that extract structured benefit information from policy documents; working knowledge of U.S. healthcare standards (ICD, CPT).

SOP Reasoner — Multi-Agent System for Automated Claims Adjudication

- Built a multi-agent system that adjudicates healthcare claims by reasoning over SOP documents, reducing adjudication cost by 80% versus the existing solution.
- Knowledge Graph Generation Agent converts SOPs into a structured decision graph of conditional business rules; Adjudication Agent traverses the graph and evaluates rules step by step.
- Produces explainable outputs (Approve / Reject / Pend) with step-by-step reasoning for transparency and auditability.

Clinical AI Expression & Semantic Search

- Built an NLP pipeline converting natural-language business requests into structured logical expressions for campaign creation, significantly reducing manual effort.
- Indexed 70,000+ healthcare concepts in Azure AI Search and improved retrieval quality with cross-encoder reranking.

Healthcare Inquiry Classification

- Developed a ModernBERT-based classifier for healthcare inquiries; addressed class imbalance with synthetic data generation and oversampling, evaluated via precision, recall, F1 and confusion matrix.

AI Research Engineer — Boltzmann Labs Pvt Ltd **Jul 2023 – Jan 2026**

LLM-Driven Agentic Workflow for RTL Verification

- Built an LLM-powered agentic workflow that extracts functional intent from IC specifications and generates coverage-aligned test plans with targeted stimulus.

- Integrated the agent into Cadence Xcelium UVM flows, reducing manual verification effort and accelerating coverage closure; benchmarked open-source and proprietary LLMs for task success and reasoning quality.

RAG & LLM Fine-tuning for Battery Chemistry

- Built a domain-specific RAG system (LangChain, Mistral API, ChromaDB) for context-aware retrieval over heterogeneous battery research data.
- Fine-tuned LLaMA-3.2-1B with LoRA on NVIDIA L40S GPUs for chemically valid crystal-structure generation (unconditional, conditional and infilling prompts).

AI Modelling of Battery Management Systems

- Implemented RNN/LSTM/GRU models in PyTorch (and from scratch) achieving 99% reliability in state-of-charge prediction.
- Applied dynamic quantization and inference optimization: 50% smaller model, 2× lower latency, <1.2% MAPE deviation — enabling real-time edge deployment.
- Designed a physics-informed surrogate model, improving interpretability of battery behaviour.

Molecular Property Prediction & MD Automation

- Built end-to-end ML/DL pipelines (preprocessing, feature selection, HPO) for molecular property prediction and similarity search.
- Engineered a GPU-accelerated, Dockerized GROMACS + OpenMM workflow deployed as a FastAPI microservice on AWS EKS — 40% faster setup, 30% higher pipeline efficiency.

Deep Learning Research Intern — Boltzmann Labs **Nov 2022 – Jun 2023**

- Contributed to Boltchem (small-molecule drug discovery) ML models; researched and implemented multi-objective optimization for simultaneous property improvement.

PROJECTS

NeuralMasteryAI — neuralmasteryai.com

- Self-built AI/ML learning platform with in-browser coding practice, progress tracking and a community leaderboard; companion YouTube channel @neuralmastery.

AI-Powered Financial Performance Reporting Agent

- LangGraph + OpenAI agent that analyzes company financials, generates narrative insights and benchmarks competitors via Tavily Search; Streamlit UI with CSV upload and feedback-driven refinement.

Regulatory QA RAG Chatbot

- GPT-4o RAG assistant with hybrid retrieval, query rewriting and memory-augmented context; containerized and deployed on Kubernetes.

Traffic Volume Prediction — Top 10, MachineHack Great Indian Hackathon 2024

- XGBoost model with context-aware feature engineering (weather, time, holidays).

EDUCATION & CERTIFICATIONS

B.Tech, Electronics & Communication Engineering — RGUKT (CGPA 9.0) **2019 – 2023**

- Gen AI Pinnacle Program, Analytics Vidhya (2024 – 2025) · Neural Networks & Deep Learning, Coursera · PyTorch for Deep Learning Bootcamp, Udemy
- Deep-ML: Top 100 rank · LeetCode · HackerRank