Abhinav Bharat Patil

abhinavp23.meta@coeptech.ac.in | +91-7972874523 | GitHub: abhinavships | LinkedIn: abhinav-patil-profile | Pune, Maharashtra, India

Summary

ML/GenAI engineer with production experience in large language models, vision-language models, agentic systems and retrieval pipelines, alongside research-grade scientific computing. Built and shipped three AI systems at Tata Motors (Digital.ai), authored a 14-module Python package for LLM weight-structure analysis at PRL–ISRO, and deployed a containerised clinical Graph-RAG service to AWS. Metallurgy and Materials Science background applied as a cross-domain differentiator across materials informatics, scientific ML and quantitative finance.

Education

2023 – 2027	**B.Tech**, Metallurgy & Materials Science Engineering — COEP Technological University, Pune	CGPA: 7.67/10
2023	**Class XII (CBSE)** — GEMS English Medium School	76.6%
2021	**Class X (MSBSHSE)** — St. Lawrence High School	93.4%

Electives Banking, Financial Services & Insurance; Python for Problem Solving

Technical Skills

Languages	Python, C++
ML / GenAI	Large Language Models, Vision-Language Models, Agentic AI, RAG, Graph-RAG, Fine-Tuning, LangChain, LangGraph, Hugging Face, PyTorch, Scikit-Learn
Data	Pandas, NumPy, ChromaDB, BGE embeddings, Postgres
Infra / Tools	Docker, Kubernetes, AWS EC2, FastAPI, HPC & SLURM, Git, OpenCV, Whisper ASR, scispaCy
Domain	Materials informatics (HEA phase classification), quantitative finance, clinical NLP

Internships

Research Intern — PRL–ISRO, GSDN Division May 2026 – Jul 2026

- Applied Random Matrix Theory to transformer weight matrices under Dr. Shivansh Verma, using spectral analysis (eigenvalue distributions) to characterise learned representations in LLM weight structures.
- Built `rmt`, a 14-module Python package that analyses the internal weight structure of large language models (LLaMA-3.1-8B, Pythia, BERT) to diagnose training quality and layer health, running at scale on **HPC/SLURM clusters**.
- Found and fixed two compounding numerical bugs in the core analysis routine that were silently corrupting every downstream result, and replaced an unstable approximation with an exact computation validated against published reference implementations.
- Used the **PAIR framework** (Propositions, Arguments, Information, Results) to systematically deconstruct and evaluate RMT/transformer research papers, sharpening rigour in reading GenAI-adjacent literature.
- Built and presented a Beamer deck synthesising RMT–transformer spectral findings for internal professor review, translating theoretical weight-space analysis into a defensible technical narrative.
- Co-designed **PRAHARI** (ISRO Bharatiya Antariksh Hackathon), a multi-horizon forecasting system using PatchTST/iTransformer — transformer-based time-series architectures — for GEO relativistic electron flux prediction.

AI/ML Intern — Tata Motors, Digital.ai Nov 2024 – May 2025

- Fine-tuned Moondream2/InternVL2 vision-language models via **LoRA** on assembly-line video to build **MOST**, a motion-analysis VLM deployed for production line monitoring, cutting manual review effort on operator motion assessment.
- Architected the **Kano Market Intelligence Agent**, a multi-agent LangGraph system orchestrating retrieval (ChromaDB + BGE embeddings) and reasoning over Anthropic's Claude API to automate competitive market/feature analysis.
- Designed the RAG pipeline for Kano end to end — embedding generation, vector indexing, and agent-to-agent handoffs — replacing manual analyst research with an autonomous multi-step retrieval-and-synthesis workflow.
- Shipped a third production VLM agent (**Ergo**) for ergonomics assessment, taking 3 GenAI systems from prototype to deployed tools used by cross-functional teams at Tata Motors.

Projects

SmartEMR — Clinical Reasoning Pipeline with Advanced Graph-RAG Dec 2025 – Present

- Clinical coding assistant that takes free-text notes and returns ranked ICD-10 codes, each linked to the exact source span that justifies it; built on scispaCy NER → ChromaDB retrieval → LLM re-ranker, with zero hardcoded medical knowledge (the ontology lives in the vector index, not in code).
- Architected as decoupled services (FastAPI request layer plus an independently scaled model inference endpoint, with Postgres + ChromaDB persistence), containerised with Docker and deployed to a live URL on AWS EC2 via Kubernetes.
- Built a human accept/reject feedback loop that accumulates labelled data, plus an evaluation harness reporting precision and recall.

High-Entropy Alloy Phase Prediction — Metallurgy & Machine Learning Aug 2025 – Present

- Developed a neural network model using Scikit-learn to predict crystal structure phases (BCC, FCC, single-phase, multiphase) in high-entropy alloys (HEAs).
- Utilised a research dataset of thermodynamic descriptors including configurational entropy, enthalpy of mixing and atomic size difference.
- Performed data preprocessing, exploratory data analysis and feature correlation studies.
- Achieved high, publishable accuracy, validating the application of machine learning to materials informatics and thermodynamic phase prediction.

Position of Responsibility

Head of Tech Development — COEP Quant Finance Club Oct 2025 – May 2026

- Lead the club's technical vertical, integrating quantitative finance with full-stack and ML engineering.
- Oversee development of web platforms and machine learning pipelines for financial modelling, data-driven trading strategies and portfolio analytics.
- Manage a multidisciplinary team of developers and analysts building scalable tools across finance, statistics and computation.
- Drive focus areas spanning automation, data engineering and predictive modelling in quantitative research.

Languages

English, Hindi, Marathi