Abhay Arora

+91-9311799994 • abhayarorasap123@gmail.com • LinkedIn • GitHub • Portfolio • LeetCode

PROFESSIONAL SUMMARY

ML Engineer architecting production agentic AI systems at Epsilon. Conference presenter (ICBAI 2025). Hands-on expertise with reasoning model alignment (GRPO/DPO) on TPUs, multi-agent orchestration (LangGraph), and AWS MLOps. Focused on end-to-end ownership from distributed training to REST API model serving.

TECHNICAL SKILLS

- **MLOps & Production Engineering**: AWS (ECS, Fargate, Lambda, API Gateway, ALB, EventBridge, ECR, CloudWatch), Docker, CI/CD, MLflow, System Design, FastAPI, Rest APIs, Model Serving
- **Generative AI & LLM Engineering**: LangGraph, Agentic AI, Model Alignment (SFT, DPO, GRPO), RAG, Vector DBs, Pydantic Schema Validation, LLM Evaluation (RAGAS)
- **Core ML & Data Science**: Python, PyTorch, JAX, SQL, Scikit-Learn, PySpark, Databricks, Pandas, NumPy, Causal Inference
- **Competencies**: End-to-End Ownership, System Design, Production Deployment, Agile Methodologies

EXPERIENCE

- **Epsilon** Bengaluru, India
  Associate Data Scientist (ML/AI Engineer) *Aug 2024 - Present*
  - Architected and deployed a suite of production-grade **agentic AI solutions**, orchestrating multi-agent workflows from R&D to **REST API** model serving. Key deliverables include an **Insight Automation Platform** cutting reporting cycles by **95%**, and a multi-agent **Audience Builder** (Text-to-SQL), rigorously validated with RAGAS to ensure **high factual correctness, safety, and a Faithfulness score of 0.9**.
  - Built and deployed a **multi-model propensity framework** on **Databricks**, using the **NLRR (New, Lapsing, Reactivated, Retained)** segmentation. Leveraged **Hyperopt** for tuning, **MLflow Feature Store** and **Model Registry** for governed deployment and drift monitoring achieving **~15% conversion lift** and **17% improved targeting precision**.
  - Served as the team's go-to expert for AI engineering, defining best practices for LLM evaluation and formally **mentoring 4 interns**, which was instrumental in achieving significant client success and process improvements.
- **Epsilon** Bengaluru, India
  Data Analyst Intern *Feb 2024 - Aug 2024*
  - Engineered a **production-grade, scaled implementation** of the **TIFUKNN** research paper in **PySpark**, delivering a **high-quality, time-aware recommendation system** for a major FMCG client, achieving a **7% lift in NDCG** over the existing baseline.
  - Developed a **projection modeling framework** for an automobile client, leveraging **advanced feature engineering** on a **10M+ user dataset** to forecast high-value customer behavior, enabling **targeted strategies** and **doubling high-ticket conversion identification**.

RESEARCH & CONFERENCES

- **12th ICBAI Conference (2025)** Oral Presentation
  "Novel Hybrid GNN-LLM Recommender System via Behavioral Semantic Profiling"
  - Presented novel methodology utilizing **Behavioral Semantic Profiling** and **InfoNCE loss**, delivering **34.5% lift in NDCG@20** over LightGCN baseline in text-scarce environments.

PROJECTS

- **Tunix: Reasoning Model Alignment (Gemma 3)** Kaggle Case Study — Model
  Tunix, JAX, GRPO, DPO, Fine-Tuning, Post-Training
  - Distilled **Chain-of-Thought reasoning** into Gemma-3-1B-IT using a novel **Metadata-Aware pipeline**, leveraging DeepSeek-V3 to generate "self-grading" synthetic datasets with hidden logic checkpoints. Engineered a memory-efficient **GRPO loop** on a single TPU v5e by implementing custom **gradient accumulation strategies** ($micro\_batch = 1$), enabling 4x parallel generation without OOM. Solved sparse reward issues by adapting the **MPO (Mixed Preference Optimization)** framework.
- **What's Good: AI Systems** Case Study & Live Link
  AI Systems Design, MLOps, RAG, AWS, CI/CD
  - Architected a distributed RAG pipeline (ECS/Lambda), reducing **p95 end-to-end latency: 17s → 6s** through asynchronous parallelization and resolving Fargate CPU throttling. Implemented **Real-Time Vector Interpolation** (NumPy) for instant persona adaptation, achieving **850ms p95 retrieval latency**. Reduced compute costs by **~60%** by migrating ingestion to Serverless and engineered a **QEMU-based CI/CD pipeline** to resolve ARM64/x86 architecture mismatches.
- **Curiocity: Multi-Agent Conversational AI**: Engineered a 0→1 multi-agent conversational AI platform. **Orchestrated** agents and served the system via **FastAPI (REST APIs)**, implementing **schema validation (Pydantic)** and structured data extraction to ensure reliable downstream integration. Fine-tuned LLMs with **Unsloth/LoRA** for specialized behaviors, enabling real-time, simultaneous multi-agent interactions.

EDUCATION

- **Vellore Institute of Technology** Chennai, India
  Bachelor of Technology - Computer Science (Spec. in AI & Robotics); CGPA: 9.09/10 *2020 - 2024*

LEADERSHIP & ACHIEVEMENTS

- **President, Zero Bugs Club, VIT Chennai** (May 2022 – May 2023): Led a **50+ member technical club**, organizing **10+ technical workshops** and **2 hackathons** that attracted over **500 participants**.
- **Databricks Certified ML Associate** and **Databricks Certified Generative AI Engineer Associate**.
- Participated in **10+ hackathons** focused on rapid prototyping and collaborative problem-solving under tight deadlines.
- Completed **Advanced LLM Agents (Ninja Tier) — Berkeley Center for Responsible, Decentralized Intelligence**.