Himadri Mishra

Senior AI Engineer

Remote, India | +91-9918889939 | himadri.jobhunt@gmail.com
linkedin.com/in/hmishra2250 | github.com/hmishra2250 | himadri.dev

SUMMARY

Senior AI Engineer (IIT-BHU, 8 yrs) specializing in agentic AI platform architecture and production LLM systems. Architected Knit 3.0's automated market research platform, replacing multi-day analyst workflows; previously led Computer Vision systems to 98% accuracy at Osmo and cut ML infrastructure costs by 10x at Epic!.

EXPERIENCE

Senior AI Engineer *Remote, India*
Knit May 2025 – April 2026

- Served as **principal architect and senior IC** for the India AI team; designed and launched the **Knit 3.0 agentic market research platform**, automating the full pipeline from raw survey data to verified insights, visualizations, and consulting-grade PPTX decks, **reducing post-fielding report turnaround from 48–72 hours to under 1 hour**.
- Engineered the **insight execution engine**: LLM writes Python analytics code, executes it in a persistent **E2B sandbox** reused across 30–50 tasks; a judge LLM verifies each result with its own sandbox code, calibrated until AI-native output **matched and surpassed the prior human-reviewed pipeline** in insight accuracy, narrative quality, and deck visual output.
- Implemented **DAG-based task orchestration with topological sorting** for maximum parallel insight execution; built the **visualization pipeline** producing 15–25 Highcharts charts per report with multi-threshold quality scoring.
- Unified all agents onto a **shared Python platform**: multi-provider LLM routing (GPT-4/5, Claude Sonnet/Opus 4.x, Gemini 2.5/3.1 Pro), YAML DAG orchestration, OpenTelemetry and Langfuse observability, pgvector-backed RAG, SSE streaming, and auto-generated REST APIs.
- Built the **data ingestion pipeline** (Knit Platform 2.0 and Qualtrics) and the **memo-to-PPTX deck pipeline**, converting research reports through intermediate representation and HTML to native PowerPoint with iterative LLM-driven visual evaluation.

Independent ML Engineer (Freelance) *Remote, India*
Stealth Product Team March 2025 – April 2025

- Built an **end-to-end text-to-vector-icon generation pipeline**, covering dataset creation, model benchmarking, LoRA fine-tuning, segmentation, and vectorization.
- Curated **18k vector icons** and generated high-quality descriptions through batch Gemini API calls; benchmarked proprietary and open-source text-to-image models including Flux and SDXL.
- Trained and evaluated thin-structure segmentation models including U-Net, U-Net++, U2-Net, and SegFormer, then built a graph-based skeleton-to-vector conversion workflow.

Career Break May 2024 – February 2025
Personal sabbatical

Senior Research Engineer *Remote, India*
Epic! for Kids February 2023 – May 2024

- **Owned full ML pipeline and infrastructure** post-layoffs across discovery, recommendation, and search; led ElasticSearch autocomplete revamp, **outperforming prior solution in 80%+ cases**.
- Cut Docker build time by **50%**, reduced Kubernetes pod usage by **100x**, slashed spot instance errors by **99%**, cutting platform cost by **10x**.
- Launched ML features (book picker, performance-based carousels) via A/B testing, improving engagement; defined and executed ML roadmap resolving high-impact infrastructure issues.

- Supported backend (PHP), frontend logging (Angular), and analytics including churn prediction.

Senior Research Engineer *Remote, India*
Tangible Play (Osmo) September 2019 – February 2023

- Acted as **Computer Vision lead** across India and US teams; set technical direction, mentored engineers, and served as primary CV point of contact for the Worksheets product.
- Raised CV accuracy in Worksheets from **93% to 98%** impacting millions of learners; built a real-time U-Net shaded-region detection model achieving **80% IoU**, deployed in Java.
- Led tracing the dots games for Pre-KG to Grade 3, **boosting engagement by 20%**; prototyped Byju's Board real-time teacher feedback system with **10fps detection**.
- Automated tagging workflows in Go and Dart, **cutting manual effort by 99%**.

Deep Learning Engineer *Bangalore, India*
Whodat (AR startup, acquired by Byju's) July 2018 – August 2019

- Built fast ORB detector in C++, **20% faster than ORB-SLAM**; researched monocular depth estimation (DeMoN, MvDepthNet). *Team transitioned to Tangible Play/Osmo post-acquisition.*

SKILLS

Generative AI and Agents: LLM Agents, Agentic Workflows, Multi-Agent Systems, Prompt Engineering, LLM Evaluation, AI Guardrails, Retrieval Augmented Generation (RAG)

AI Frameworks and APIs: LangChain, LlamaIndex, Hugging Face, Sentence Transformers, BERT, OpenAI API (GPT-4/GPT-5), Anthropic API (Claude Sonnet/Opus 4.0–4.6), Google Gemini 2.5/3.1 Pro

Machine Learning and Domains: PyTorch, TensorFlow, Computer Vision, Natural Language Processing (NLP), Survey Analytics, Market Research, MLOps, Statistics, Optimization

Programming Languages: Python, TypeScript, C++, Java, Go, Dart, PHP

Backend, Data and Platform: FastAPI, REST APIs, SSE, PostgreSQL (pgvector), Redis, S3, MinIO, SQL, PySpark, ElasticSearch, Highcharts, OpenTelemetry, Langfuse, E2B, Kubernetes, Docker, Git, Google Cloud, CI/CD

Databases: MySQL, MongoDB, SQLite

Web and Frameworks: Django, Angular, Next.js, React

PROJECTS AND AWARDS

Kaggle: FIDE and Google Efficient Chess AI Challenge (2025): Secured **top 6% globally**, Rank 60 out of 1120.

Research Intern, UC Berkeley (2017): Trained neural programmer-interpreters using MuJoCo under Professor Dawn Song. *SN Bose Scholar Program Awardee*, 2017.

Software Engineering Intern, Microsoft (2016): Developed dialog-based chatbots for the Incentive Compensation team.

Earlier competitions: Top 8% globally, Kaggle Allstate Claims Severity (2017); Honorary mention ACM ICPC Amritapuri Regional (2017); Honorable mention HP Think-A-Thon (2016).

EDUCATION

Indian Institute of Technology (IIT-BHU), Varanasi 2013 – 2018
Integrated Dual Degree in Computer Science and Engineering
GPA: **9.28/10**