# Ankit Chauhan

(317) 491-7832 | ankichau.1718@gmail.com | LinkedIn | Portfolio

## SUMMARY

AI Research Engineer specializing in Natural Language Processing (NLP) and agentic AI systems. Hands-on designing multi-model agent architectures - model routing, tool-use, retrieval, and prompt orchestration - on serverless AWS (Lambda, DynamoDB, API Gateway) using Python and TypeScript. Proven track record fine-tuning LLMs (LoRA/QLoRA, RAG) and shipping production LLM services, with over 2.5 years of applied/research ML experience on top of approx. 4 years of Azure cloud engineering (Capgemini + Wipro).

## EDUCATION

**Indiana University, Indianapolis** **Aug 2023 - May 2025**
*Master's, Applied Data Science (GPA: 3.87/4)* *Indianapolis, IN*

- **Coursework:** Deep Learning, Computer Vision, Predictive Analytics, Big Data, Project Management

**University of Mumbai** **May 2020 - May 2023**
*Bachelor's, Information Technology (GPA: 9.08/10)* *Mumbai, India*

- **Coursework:** Python, Data Warehousing, Operating Systems, Java, OOPM, Data Structures, AI

## SKILLS

**Languages**: Python, SQL, TypeScript, Bash, Go, R, C++
**AI/ML**: PyTorch, SpaCy, TensorFlow, Scikit-learn, LightGBM, statsforecast, Pandas, NumPy, Time-Series Forecasting, Information Extraction, NER, Knowledge Graphs
**LLM Workflows**: HuggingFace Transformers, Fine-tuning, LoRA/QLoRA, RAG, LangChain, RAGAS, MLFlow
**Cloud/MLOps**: Docker, GitHub CI/CD, Amazon Bedrock, AWS Lambda, AWS SAM (IaC), S3, EventBridge, CloudWatch, Amazon ECR, API Gateway, FastAPI, RESTful API Development, React, Cytoscape.js
**Databases**: MySQL, PostgreSQL, MongoDB, DynamoDB, Elasticsearch, Qdrant, Neo4j, Cypher
**Research**: Statistical Analysis, Ablation Studies, Experimental Design, A/B Testing
**Certifications**: Microsoft Certified: Azure AI Fundamentals (AI-900), DeepLearning.AI Deep Learning Specialization, IBM Data Science Professional Certificate

## EXPERIENCE

**Indiana University Indianapolis**
*AI Research Engineer* *Indianapolis, IN · Aug 2025 - Present*

- Architected a multi-model agentic pipeline on Amazon Bedrock Agents (Lambda, DynamoDB) that routes documents via a Claude Haiku classifier between Claude Sonnet and a fine-tuned Mistral-7B, cutting per-token cost approx. 20 times through task-based model arbitration.
- Raised structured-extraction accuracy from 0.32 to 0.78 Triple-F1 by fine-tuning Mistral-7B with LoRA + RAG over the project ontology.
- Turned unstructured community-research transcripts into a queryable Neo4j knowledge graph (8.3k human-reviewed records), serving as a structured knowledge base for downstream agent retrieval.
- Developed a human-in-the-loop review UI (React, Cytoscape.js, FastAPI) that captures every researcher correction as an audit trail and model training data, validated at 0.82 Cohen's Kappa on a double-annotated sample.

*Research Assistant* *Indianapolis, IN · Sep 2023 - May 2025*

- Curated a 7,200-item instruction-tuning corpus reliable enough to train on (0.88 Cohen's kappa inter-annotator agreement) for an NSF-funded educational AI project.
- Improved an educational AI tutor's pedagogical alignment 14% and cut hallucinations 8% (versus baseline, held-out eval) by fine-tuning Llama-3.2 with QLoRA and DPO.
- Taught 30+ graduate students deep learning as TA for Fall 2024 H518: labs, tutorials, office hours, and lesson plans on GANs, CNNs, LSTMs, Transformers, and RL.

**Capgemini**
*Cloud Engineer* *Mumbai, India · Apr 2020 - Sep 2022*

- Held 23,000+ users within SLA through an Azure tenant split by owning end-to-end migration of identity (Entra ID), mailboxes, and legacy apps onto provisioned VMs, vNets, storage accounts. Recognized in top 5% of employees (FY 2022).
- Saved approx. 15% in annual Microsoft 365 and Azure spend across the tenant by profiling SKU and resource utilization and surfacing idle workloads and underused licenses through recurring Azure CLI reports.

- Ramped 57% of new hires to production-ready Azure operations (portal, RBAC, governance) by running hands-on enablement and authoring 20+ runbooks across subscription governance, VM and networking troubleshooting, and identity.

**Wipro Limited**
*Cloud Engineer* *Navi Mumbai, India · Jul 2018 - Mar 2020*

- Pinpointed recurring failure patterns and root causes across 1,500+ Azure incidents by mining infrastructure-performance data, sustaining 85%+ CSAT on resolutions spanning IAM, VMs, Storage, VNets, and Azure AD.

## PROJECTS

**Demand Forecasting Copilot: Intermittent-Demand Forecasting with an Agentic Layer** **May 2026**
*Personal project · GitHub*

- Built a regime-aware forecasting router using ADI/CV$^2$ for 14,370 Walmart (M5) series (22M+ observations) that routes each demand regime to its best model (TSB or a global LightGBM), cutting held-out WAPE by approx. 25% over a naive baseline, beating all nine benchmarked models (Croston variants and a tuned moving average).
- Deployed it as an MLflow-versioned model on AWS Lambda with automated drift monitoring, and layered on a Bedrock copilot, combining tool use over the live forecast API with RAG over a retail knowledge base.

**CultureEval: Quantifying Cultural Alignment in LLMs** **Aug 2024 - May 2025**
*Indiana University Indianapolis · GitHub* *Indianapolis, IN*

- Reduced survey responses from 97k people across 96 sociocultural indicators to five latent cultural dimensions via PCA, building the framework in Python (pandas, NumPy, scikit-learn).
- Revealed that Llama-2 13B, Gemma 3 12B, and Phi-4 underestimate Religious-Traditional values for non-Western profiles (Cohen's d: 0.89–1.17), measured via Tucker's Congruence Coefficient and Cohen's d.

**Analyzing Chart-to-Text with CNN–RNN and Vision-Language Models** **Aug 2024 - Dec 2024**
*Indiana University Indianapolis* *Indianapolis, IN*

- Benchmarked CNN–RNN captioners against modern vision-language models on 27k+ Statista charts (preprocessing images and extracting visual features) to generate natural-language summaries.
- Nearly tripled caption quality (BLEU-4 from 0.18 to 0.50) by swapping ResNet-50 for EfficientNet-B2 and integrating a dual-LSTM decoder with coverage attention in PyTorch.

## PUBLICATIONS

- **Mod-Guide: An LLM-based Content Moderation Feedback System.** *ACM COMPASS 2026 (Accepted).*
  Built a RAG-grounded, persona-prompted LLM moderation system over a community-sourced corpus that measurably outperforms off-the-shelf GPT-4 on contextual accuracy.
- **Benchmarking LLMs for Pairwise Causal Discovery in Biomedical and Multi-Domain Contexts.** *2025 IEEE International Conference on Big Data.*
  Benchmarked 13 open-source LLMs on causal detection and extraction across 12 datasets and 5 prompting strategies to surface where they rely on explicit markers and fail on implicit, inter-sentential text.