Kanchan Bhale

kvb2117@columbia.edu | +1(646)406-3547 | LinkedIn | Portfolio | New York, NY | github.com/Kanchanbhale| Certificates

Education

Columbia University : Columbia Engineering and Graduate School — New York, NY

M.S. in Data Science — *Aug. 2025 – Dec 2026*

Key Courses : Systems Design, LLM Based Generative AI, Algorithms, Machine Learning, Agentic AI, ML for Quantitative Finance

Engineering Graduate Student Council : Selected as Data Science **Department Representative**

Teaching Assistant - Agentic AI, Graduate Researcher - Columbia Data Agents Process Lab (DAPLab) : LangChain, LangGraph

MKSSS's Cummins College of Engineering for Women — Pune, India

B.Tech. in Electronics and Telecommunications Engineering — *Aug. 2018 – May 2022*

Key Courses : Probability & Statistical Inference, Data Structures, Artificial Intelligence, Natural Language Processing, Databases

Industry Experience

Ask2.AI — Aug 2025 – Jan 2026

Quantitative Student Researcher

- Built a HMM-based detection pipeline (R2-RD) to dynamically update regimes over time, improved robustness by more than 80% using percentile-based expansion rules. Designed regime-aware explainability by training decision trees. Computed SHAP feature attributions, achieved 30–45% higher cross-regime interpretability in predicting next-period macroeconomic regime transitions.

The Boeing Company — Aug 2022 – Aug 2025

Data Scientist II

- **Boeing Conversational AI : Code Assistant Multi-Agent System**- One of core engineers who built Boeing's first enterprise LLM code-assistant by customizing Continue.dev with internal secure and private Conversational AI APIs, reducing code-authoring latency by 35% and accelerating developer adoption across all Boeing teams.
- Designed and deployed TypeScript-based telemetry and auto-correction pipelines to track model performance, feature usage, and failure patterns, driving a 25% boost in suggestion relevance and a 30% drop in incomplete or failed completions.
- **End-to-End ChatBot RAG Pipeline** - Built a Retrieval-Augmented Generation (RAG) pipeline with ChromaDB, performed hyper-parameter tuning of a Sentence Transformer, improving retrieval accuracy by 25% and cutting document search time by 40%, led to estimated $2M+ annual productivity savings across Boeing engineering teams.
- Developed Flask APIs for user creation, filtering, and personalized recommendations, enabling seamless integration by collaborating with front-end engineers for robust API performance, reducing query latency by $\tilde{3}$0%.
- **Computer Vision for Mitigating Foreign Object Debris in Aircraft Assembly** - Spearheaded a vision agent project; built a unified evaluation script to benchmark the accuracy of the entire system, improved detection accuracy by 30%. Trained, tested, and deployed real-time models on HPC infrastructure; exposed scalable APIs for model access, cutting manual inspection effort by 40%
- **Published** 3 Invention Disclosures in Multi-Agent Systems, LLMs for Security and Risk Management in Boeing Technical Journal
- One of the 38 globally selected members for a very competitive AI rotation program across Boeing Enterprise

AI Engineer Intern — May 2021 – July 2021

- Engineered and deployed a real-time ML + NLP inference service (Python, FastAPI) combining factory telemetry with API signals and productionized data pipelines improving cycle-time prediction accuracy by 25% enabling automated production line adjustments.

Projects-[portfolio]

1. (Winner) Snapdragon-Qualcomm Hackathon - Edge CV + Multi-Agent RAG + Multi-LLM API [code]

- Built a privacy-first edge ML pipeline with INT8 YOLOv8 (ONNX Runtime) on mobile, publishing typed event streams under sub-100ms, offline latency constraints. Designed a control-plane multi-agent reasoning orchestration layer using heterogeneous LLMs (GPT-4o-mini, Gemini Flash, Claude Sonnet) to drive HITL-gated triage and escalation for safety-critical decision workflows.

2. Amazon Bedrock + AWS Lambda Powered Autonomous LLM-Orchestrated DevOps System [code]

- Built a multi-agent Bedrock–Lambda DevOps system that auto-generated & validated code patches with more than 90% diff-accuracy and less than 3s serverless orchestration latency across GitHub/Jira workflows. Engineered an RAG-enhanced LLM CI/CD loop using vectorized context + multi-step tool-calling, enabling 70%+ automation of debugging, commits, and deployment tasks.

3.(Top 18/600) NVIDIA x Dell Hackathon - Vision Agent AI Fall-Risk Detection [code]

- Deployed real-time YOLOv8 pipeline using pose dynamics + temporal inference, optimized on Dell GB10 large-compute hardware.
- Integrated NVIDIA Nemotron-4 340B for agentic reasoning over CV outputs, enabling autonomous SOS decisioning within web app

4. Trainium-Optimized Multi-Modal Agentic Task Automation System (SLMs + CrewAI) [code]

- Built a Trainium-accelerated SLM inference + event-driven context router (Python, Triton, CrewAI, AWS STT) for real-time multi-modal task automation, achieving 3x throughput gains, more than 85% intent/slot-filling accuracy, and a 60% lift in actionable task execution across voice, email, notes, and calendar streams.

Technical Skills and Frameworks

Programming & Development: Python (NumPy, pandas), FastAPI, C++ , Flask, Streamlit, REST APIs, SQL, OOP, Git.

AI/ML & LLMs: Hugging Face, Seaborn, Matplotlib, Pillow, OpenCV, NLTK, CI/CD Pipelines, GitHub Actions, Airflow.

Data Infrastructure: Vector Databases (FAISS), Ollama, Agent Orchestration, Model Evaluation, PySpark.

Cloud & MLOps: AWS (Lambda, SageMaker, S3, Aurora, API Gateway, SQS, CloudFront), Azure, GCP, Terraform (IaC), Docker.

Research Publication

1. **Technical Analysis of Artificial Intelligence Assisted Swarm CubeSats for Active Debris Removal in LEO**. IAC-21,A6,5,x65873, 72nd International Astronautical Congress(IAC). [link]
2. **Multilingual Large Language Model Transformer for Indigenous Languages**. 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025), Vienna, Austria July 27 to August 1st, 2025 . [link]
3. **Ethical Considerations When Deploying ML Systems**. Forty-Second International Conference on Machine Learning (ICML 2025), 13-19 July 2025. [under review]