# Ameya Ranade

413-479-9346 | aranade@umass.edu | ameya-ranade | ameyaranadee

## Education

**University of Massachusetts Amherst** — Aug. 2023 – May 2025
*Master of Science in Computer Science;* ***CGPA 3.92/4*** — *Amherst, MA, USA*
*Relevant coursework: Advanced NLP, Systems for Data Science, Systems for Deep Learning, Distributed and Operating Systems*

**University of Mumbai (S.P.I.T.)** — Aug. 2019 – May 2023
*Bachelor of Technology in Information Technology, Minor in Management;* ***CGPA 3.72/4*** — *Mumbai, MH, India*
*Relevant coursework: Data Structures, Algorithms, Distributed Computing, Databases, Computer Networks*

## Experience

**Intellect Design Arena** — July 2025 – Present
*Software Engineer* — *New Brunswick, NJ, USA*

- Engineered **4+** core features for **Xponent**, a commercial insurance underwriting workbench, using Java and Spring Boot to automate compliance checks, risk analysis, and premium calculations.
- Built admin configuration screens using Typescript and Angular, implementing session-based caching to eliminate redundant database calls, slashing UI load times by **60%.**

**Apple** — Feb 2025 – June 2025
*Graduate Student Researcher Extern* — *Amherst, MA, USA*

- Developed few-shot entity linking system using LLMs for detecting and linking salient entities in web articles that outperforms state-of-the-art methods by **+20.9** F1, enabling more accurate and scalable entity-centric indexing for Safari. (Paper)
- Designed few-shot LLaMA-3-based workflow jointly optimizing entity recognition and salience prediction, outperforming a supervised pipeline by **+16.2 F1** on out-of-domain datasets.

**Priro Systems** — June 2024 – Dec 2024
*Software Engineer* — *Dallas, TX, USA*

- Built **10+** end-to-end features for SpecGenie, an AI-powered EPC workflow automation product for industrial engineering specifications, reducing processing time from **weeks to under 4 hours**.
- Architected document extraction leveraging AWS Textract and LLMs to convert unstructured piping specifications into structured enterprise data with **98% accuracy**, processing **1000+** components in **under 2 minutes.**
- Designed compound AI systems combining LLM orchestration and rule-based validation to generate **35+** bulk-loadable specification and catalog Excel sheets, eliminating **95%** of manual data entry.
- Developed dashboards with React and Material UI, integrating **15+** RESTful endpoints and CRUD support for specification lifecycle management and revision tracking.

**Center for Machine Intelligence and Data Science (CMInDS)** — Jan 2022 – July 2022
*Machine Learning Research Intern* — *Mumbai, MH, India*

- **Automatic Short Answer Grading.** Developed autograder using stacked embeddings and ensemble models, beating state-of-the-art results by **13%**. Implemented custom spell correction, cosine similarity variants, and per-question training to improve semantic matching accuracy by **15%.**

**LokaVidya Technologies** — July 2021 – Jan 2022
*Software Engineer Intern (Backend)* — *Mumbai, MH, India*

- Developed **LokaVidya Meet**, a video conferencing platform using Node.js and BigBlueButton APIs to deliver **6+** meeting features, supporting secure virtual meetings across educational institutions.
- Migrated RBAC from in-memory storage to Redis for session and role caching, reducing access latency by **87%** and offloading Azure SQL queries on subsequent requests.

## Technical Skills

**Languages**: Python, Java, C/C++, JavaScript, TypeScript
**Full Stack:** Django, Flask, FastAPI, Node.js, Express, Spring boot, Next.js, React, Angular, REST, Postgres, MongoDB, SQL
**Data Science & AI/ML:** PyTorch, TensorFlow, NumPy, Sklearn, Matplotlib, Seaborn, SpaCy, Huggingface, Pydantic, LangChain, Unsloth, vLLM, ChromaDB, Pinecone
**DevOps:** Docker, Docker Compose, AWS, Azure, Kubernetes, Redis, CI/CD, Kibana, Grafana, Camunda

## Projects

**Distributed Stock Bazaar** | *Python, Flask, Docker, AWS, REST APIs*

- Designed and deployed fault-tolerant microservices for simulated stock trading on AWS EC2 using Flask RESTful APIs and Docker, handling **500+** concurrent client requests. Implemented LRU caching with server-push invalidation, **reducing query latency by 40%** and engineered leader-election mechanisms across replicas to ensure zero data loss.

**ArxivApp** | *FastAPI, Next.js, Tailwind*

- Built a research-paper QA app with batch ingestion, section-aware text chunking for context preservation, and semantic search across arXiv papers.

**Reflective Prompting: Self-Refine RAG for Piazza QA** | *PEFT, Generative AI, Unsloth, vLLM, FAISS, LangChain*

- Built an iterative RAG architecture using FAISS for vector search and Llama3-8B (via Ollama) for local inference on custom QA dataset to eliminate model bias and improve accuracy by **8%** over baseline. (Code)