

AMEYA RANADE

(413) 479-9346 | ranadeamr@gmail.com | ameyaranade.com | linkedin.com/ameya-ranade

EDUCATION

University of Massachusetts Amherst

Master of Science in Computer Science (GPA: 3.92 / 4.00)

May 2025

Amherst, MA

University of Mumbai (Sardar Patel Institute of Technology)

Bachelor of Technology in Information Technology, Minor in Management (GPA: 3.72 / 4.00)

May 2023

Mumbai, India

WORK EXPERIENCE

Intellect Design Arena

Software Engineer

Jul 2025 – Present

Piscataway, NJ

- Delivered feature implementations, client customizations, and production bug fixes for **Magic Submission**, a multi-tenant broker submission intake and extraction platform processing **100k+** docs/month across **9+ commercial insurance lines**.
- Built Python automation tool to flag and reprocess stale submissions, **eliminating SLA breaches** across all tenants.
- Developed internal control center visualizing orchestration flows chaining **50+** services with per-step metadata, surfacing errors and **cutting RCA time** for production issues; adopted as default debugging tool by support team.
- Implemented event-driven audit log for **Xponent**, a commercial insurance underwriting platform, with Spring Boot and AWS SQS to capture field-level submission changes, **shrinking audit review time by 65%**.
- Streamlined broker follow-ups for incomplete submissions using Java, Spring Boot, and LLM-based structured extraction from email and documents, reducing manual coordination and **accelerating Time-to-Quote**.

Apple

Graduate Student Research Extern

Feb 2025 – Jul 2025

Remote

- Developed **ELAPSE** for Apple's Siri & Search team, a unified model that jointly performs mention detection, entity disambiguation, and salience prediction in a single forward pass, enabling entity-centric indexing and search at web scale.
- Built Refined-Salience, a fine-tuned adaptation of **ReFinED** for salient entity linking, outperforming prior state-of-the-art (SWAT) by **+34.4 F1** across in-domain and cross-domain settings.
- Engineered and benchmarked vLLM inference pipelines for Qwen3.5 (2B–9B) on NVIDIA H100 GPUs, configuring FlashAttention-3 and automatic batching for LLM-based salience prediction; achieved up to **25.9 docs/s** throughput.

Autire Technologies

Software Engineer

Jun 2024 – Dec 2024

Dallas, TX

- Built SpecGenie, an AI platform that automates piping specification digitization for industrial plant design, compressing multi-week engineering workflows to under 4 hours; demoed to Fortune 500 energy majors.
- Owned document extraction pipeline from ingestion to validation using Python, AWS Textract, LLMs, and Pydantic to convert unstructured engineering specifications into structured data with **98%+ field-level accuracy** across diverse formats.
- Engineered output generation layer with Celery batch processing, producing structured Excel outputs compatible with 5 major plant design platforms and **eliminating 95% of manual data entry** across spec and catalog generation.
- Designed human-in-the-loop review interface in React and Material UI for engineers to inspect, correct, and approve component extractions before downstream output generation.

LokaVidya Technologies

Software Engineer Intern

Jul 2021 – Jan 2022

Mumbai, India

- Built multi-tenant RBAC and SSO authentication for **IV Meet**, integrating core organization APIs to resolve permissions across organizations and preserve meeting context through deep-link authentication.
- Integrated BigBlueButton APIs for meeting provisioning, role-specific access, and organization-scoped recording retrieval using Node.js and Azure SQL, supporting live deployment across Ekal Vidyalaya's network of **56,000+** rural teachers.

SELECTED PROJECTS

Sift | TypeScript, Next.js, Python, FastAPI, Gemini, pgvector

- Developed a multimodal artwork search tool over **200,000+** digitized works from Harvard Art Museums collection, embedding and indexing the full catalog for text-based retrieval.

Pravas (pravas.xyz) | Typescript, Next.js, Supabase, Mapbox GL, Sarvam AI, Vercel

- Built travel journaling app for capturing and reliving trips through voice journals, photos, and location-based memories, integrating Sarvam AI for multilingual voice support.

TECHNICAL SKILLS

Languages: Python, Java, JavaScript, TypeScript, C++, SQL

AI Engineering: PyTorch, vLLM, CUDA, Pydantic, LLM APIs, RAG, vector databases (Pinecone, pgvector), tool calling, agent orchestration, fine-tuning, prompt engineering, evals

Full Stack: Django, FastAPI, Spring Boot, Node.js, React, Next.js, Angular, Svelte, PostgreSQL, MongoDB, REST

Cloud & DevOps: AWS, GCP, Docker, Kubernetes, Grafana, Datadog