INAN TAINWALA

New York, NY
(919) 695-6234
inantainwala@gmail.com

TECHNOLOGIES AND AWARDS

Patent: 20250110979 Distributed Orchestration of Natural Language Tasks Using a Generative Machine Learning Model
Proficiencies: Java, Python, AWS, Databases, Distributed Systems, ML, LLMs, CI/CD, Engineering Operations
Exposure: JavaScript, TypeScript, Express, AngularJS, React, D3.js, Play Framework

---

PROFESSIONAL EXPERIENCE

Amazon Q Business | Senior Software Development Engineer   2023 - Present

- Led a tiger team of 8 engineers to build an LLM accuracy test framework – a novel, modular and scalable solution in a nascent GenAI Evaluation space – which was crucial to launch Amazon Q Business at re:Invent 2023. It is currently adopted by a 400 person organization to test, measure and report key accuracy metrics to Amazon S-team leaders.
- Spearheaded the 'Recursive Retrieval' project that improved answer "Correctness" by 6-10%. Worked with customers and SAs to identify incompleteness issues and collaborated with science teams on proposals. Delivered an innovative 'passage-stitching' MVP that reduced time-to-production by 85%, which helped retain 6 of the service's largest enterprise customers under tight deadlines.
- Formed a new "Accuracy" team to drive quality improvement for the newly launched Q Business service. Scaled this team by developing an end to end "Quality Program" that aligned support, product, data science, engineering and ops teams to successfully triage, prioritize, experiment, test and resolve customer issues. This effort successfully resolved 75% of accuracy tickets, reduced operational load and earned customer trust.
- Designed, developed and launched the Hallucination Mitigation feature in close collaboration with science and product teams. Customers reported up to a 20% improvement in 'Faithfulness' and an increase in trust in the service's responses. This in turn improved customer enterprise adoption of the service.

Amazon Kendra | Software Development Engineer II   2020 - 2023

- Coordinated an Inferentia fleet migration effort to optimize Kendra's ML infrastructure. Diagnosed and reported complex *neuron-rtd* kernel issues through technical deep dives and partnership with ML infra teams, yielding significant CapEx savings for the organization and overall chip upgrades for AWS.
- Built a distributed Consistent Hashing algorithm using virtual nodes, as part of a multi-tenancy project to support Kendra's ISV customers to scale. Unlocked an order-of-magnitude improvement in scalability by enabling k/n vs. n redistribution of items during resharding operations.
- Led a cross-functional team to expand Kendra's natural language capabilities for 7 new languages. Integrated linguistic morphological NLP libraries for non-English languages that unblocked expansion for global customers like Pfizer, Bank of Itaú, and Barclays.

Amazon Prime Pantry | Software Development Engineer I   2017 - 2020

- Implemented a Tier-1 offer filtering service for Amazon Pantry that protected against site-wide outages, translated to an availability of ~99.999% and backed ~4M requests per minute.

EDUCATION

B.S. Computer Science + B.S.E Electrical and Computer Engineering *Duke University, Durham*   2012 - 2016
Amazon Machine Learning University *Amazon, Seattle*   2018 - 2020