

Anurag Akkiraju

Software Development Engineer II

Seattle, WA | [LinkedIn](#) | [GitHub](#)

AI systems engineer building long-horizon agent runtimes and LLM platforms at Amazon, with expertise in prompt caching, multi-agent pipelines, and evaluation harnesses.

EXPERIENCE

Software Development Engineer II

Aug 2022 — Jul 2024

Amazon

Seattle, WA

Building next-generation personal AI agents, owning the runtime for long-horizon durable work and the proactive layer that listens for real-world changes. Led the agent's prompt-caching program, improving cache hit rate from 45.6% to 83.0% while reducing latency by 57%. Led the Alexa+ notification domain rebuild, designing the model interface and evaluation harness for ~90M customers. Designed a platform primitive for deterministic service actions and contributed to the real-time proactive content system, which achieved 26x engagement improvement.

Machine Learning Research Intern

Jun 2021 — Aug 2021

Lawrence Livermore National Laboratory

Livermore, CA

Designed and trained an LSTM in PyTorch for sequence labeling over raw binary data, classifying non-code byte regions in PE and ELF binaries at 95%+ accuracy. Built an approximate nearest-neighbor retrieval system for software origin tracing using NMSLIB, returning k-nearest neighbors for newly ingested binaries against an enterprise corpus. Extended the work through spring 2022.

EDUCATION

University of Florida, Herbert Wertheim College of Engineering

B.S. Computer Science, Minor in Electrical Engineering

PROJECTS

Decision layer between robot planner and policy, Calibrated Decisions at the Human–Robot Boundary

Investigated the decision layer between a robot's planner and its policy (act, ask, or hand off to a person). Tested calibrated models against frozen rules and an oracle on four simulated bodies across 160+ experiments. Distilled a 421M-parameter on-device model from a cloud judge, maintaining 88.3% accuracy versus 87.9% for the cloud version at 90ms latency with calibration intact. Identified a failure mode in fleet learning where retraining on operator takeovers can make models confidently wrong outside corrected regions.

SKILLS

long-horizon agent runtimes · agent memory · multi-agent LLM pipelines · prompt caching · PyTorch · reinforcement learning · calibration · hybrid FTS5/dense retrieval · evaluation-harness design · instruction specs · frozen baselines · regression testing · non-inferiority testing · failure-mode analysis · Python · Java · C++ · TypeScript · SQL · Bash · AWS · Docker · Git · CI/CD · distributed systems · event-driven pipelines