THARUN BETHI

Staff Machine Learning Engineer at Apple

Seattle, WA | tharunreddy.bethi@gmail.com | Website | LinkedIn | GitHub

SUMMARY

Staff Machine Learning Engineer with 10+ years building production AI, platform, and security systems at Apple, Meta, and Amazon. Recent work focuses on LLM evaluation, grounding, retrieval, and agent systems for Siri and Apple Intelligence, with end-to-end ownership across model workflows, backend services, developer tooling, and infrastructure. Earlier work spans real-time applied ML for communications at Meta and large-scale network security and deployment systems at Amazon.

EXPERIENCE

Apple | Seattle

Staff Machine Learning Engineer | Nov 2022 – Present

- Architected and launched an LLM-powered scenario grounding platform for Siri and Apple Intelligence evaluation, processing 10,000+ scenarios per release across 8 locales and raising grounding coherence from 71% to 95% across 51 production releases.
- Authored a spec-driven AI development methodology adopted by 15 contributors and built 13 Claude Code skills, 4 commands, and 2 autonomous agents, increasing merged PR throughput 2.7x from 76 to 206 per month and supporting delivery of 150+ specs across 14 work areas.
- Replaced a 6-step sequential prompting pipeline with a LangGraph multi-agent architecture composed of 5 specialized agents and an 18-node StateGraph, adding context summarization, prompt caching, and auto-retry to improve coherence by 14+ percentage points and enable feedback-driven self-correction.
- Built hybrid retrieval across 26 entity domains using vector search, temporal queries, reciprocal rank fusion, content-hash caching, and AST-based SQL validation to improve grounding coverage while preserving reliability and security.
- Delivered the platform end-to-end across application and infrastructure layers: FastAPI backend with SSE streaming, React grounding studio and trace viewer, and Kubernetes/Terraform infrastructure with HSM-backed encryption, mTLS, Private CA pools, and multi-strategy authentication.
- Built deterministic entity hallucination detection that validates generated references against persona data and triggers automatic regrounding with structured feedback, eliminating fabricated entities from evaluation datasets.
- Led development of a synthetic persona generation platform for Siri evaluation, generating realistic test users across 33+ iOS apps in 29 locales and automating creation of 500 personas per day for Apple Intelligence test infrastructure.
- Built a LangGraph and FAISS-based radar triage agent over 3,267 historical radars, automating classification across 1,000+ radars and achieving the highest radar closure rate in the organization.

Meta | Seattle

Senior Software Engineer | May 2022 – Nov 2022

- Led a greenfield applied ML effort to improve WebRTC resilience across Facebook, Instagram, and Messenger.
- Designed an LSTM time-series model to predict network degradation and dynamically tune forward error correction and proactive packet redundancy, reducing effective packet loss by 3% and improving packet loss concealment by 1.8% under lossy network conditions.

Amazon | Seattle

Senior Software Engineer | Jul 2014 – May 2022

- Led design and rollout of a security configuration deployment service supporting 1,000+ concurrent deployments, reducing network configuration rollout time from several weeks to under 24 hours; by Apr 2022, more than 55% of security configuration deployments used the service.
- Built Amazon's first centralized network port scanning capability for 100,000+ internet-facing devices across 20+ regions, reducing time to vulnerability detection from no global visibility to under 1 hour.
- Re-architected critical legacy services by eliminating single points of failure and migrating core components to AWS-managed services, improving availability and strengthening operational resilience, and reducing host patching cycles from over 1 week to under 1 day.
- Founding member of the AWS Network Security Services organization; helped grow the team from 4 to 50+ people, conducted 200+ interviews, and mentored 15+ peers and interns.
- Received a rare Outstanding annual performance rating in 2016 in recognition of high-impact delivery in network security engineering.

EDUCATION

University of Pennsylvania

Master of Science in Engineering | 2012 – 2014

Birla Institute of Technology and Science, Pilani

Bachelor of Technology | 2008 – 2012

PUBLICATIONS

- Co-author, Context Tuning for Retrieval Augmented Generation, EACL 2024. Led data generation and contributed experimentation for retrieval-augmented generation workflows.

TECHNICAL RANGE

- Applied AI: LLM systems, agentic workflows, RAG, hybrid retrieval, evaluation, hallucination detection, synthetic data generation

- Languages: Python, TypeScript, Scala, Rust
- Platform: FastAPI, React, Bun, Kubernetes, Terraform, Docker
- Data & Search: FAISS, DuckDB, LanceDB, SQLite, vector search
- Systems: Distributed systems, platform engineering, security engineering
- Cloud: AWS, GCP