Yanpeng Qi

Bellevue, WA | yanpengqi.com | (267) 909-3678 | qyanpeng1995@gmail.com

Technical Skills

Languages: Java, Python, TypeScript, JavaScript, SQL, HTML/CSS, Shell, C++
Technologies: React, React Native, Redux, Next.js, Spring Boot, Node.js, FastAPI, PostgreSQL, MongoDB, DynamoDB, Redis, Docker, Cypress, AWS (Lambda, API Gateway, CloudWatch, ECS, Fargate, Bedrock), Claude API, RAG, pgvector, ChromaDB, MCP

Professional Experience

Software Development Engineer — April 2021 – Present
Amazon — *Seattle, WA*

Parent Dashboard

- Directed Q4 peak readiness for 1M+ users at ∼1,000 TPS. Built a Python pipeline to auto-extract endpoints, model growth curves, and forecast capacity, cutting manual analysis from 4 to 2 weeks and de-risking holiday scaling.
- Led platform migration for 200K+ users via percentage-based rollouts and URL-driven state handoffs. Preserved session continuity across stacks without full-page reloads, achieving zero downtime and a 40% reduction in page-load latency.
- Re-architected a legacy AUI app to React/Spring Boot. Designed a config-driven routing engine for OOBE flows, enabling 10+ partner teams to onboard device experiences via JSON, cutting onboarding time from 2 weeks to 3 days.
- Mentored 2 interns and built an AI knowledge assistant using a RAG pipeline and custom MCP server, centralizing fragmented wikis into a single interface to significantly reduce ramp-up friction for new engineers.

IAM Organization

- Built an identity-aware microservice on AWS Fargate to automate organization-membership validation. Enforced fine-grained IAM boundaries and tenant isolation, launching a scalable, auditable control point across all 31 AWS regions within 3 months.
- Integrated an AI code review workflow into CI/CD via AWS Lambda and Bedrock LLMs. Designed severity-tiered prompts to distinguish critical security issues from suggestions, reducing manual reviews by 25% and duplicate tickets by 20%.
- Validated an asynchronous job-orchestration system for large-scale metadata synchronization. Simulated 10,000+ daily jobs to verify resilient flows and configured custom CloudWatch alerting to proactively surface authentication anomalies.

Projects

Admitly — Agentic AI Assistant for Graduate Admissions — 2025

- Orchestrated 16 AI ops across 4 pipelines (12-step Deep Research, 5-stage essay coach); validated via LLM-as-judge, achieving 78% pass rate on 10 SOP prompts.
- Built a 92% precise RAG pipeline (pgvector BM25+dense+RRF, .edu scraping). Designed an Opus/Sonnet/Haiku model router, cutting per-pipeline LLM costs by 54%.

FinSight — AI Financial Intelligence Agent — 2024

- Engineered a Claude agent chaining 5 tools (news, SEC, prices, technicals, sentiment) to generate cited investment signals in <3s latency across 500+ articles/day.
- Built news clustering (0.92 cosine threshold) with velocity tracking, cutting false alerts by 45%. Added a few-shot sentiment classifier (87% acc, 79% precision).

Education

University of Pennsylvania — Philadelphia, PA
Master of Science in Computer and Information Technology — *Aug 2018 – Dec 2020*

Sun Yat-Sen University — Guangzhou, China
Bachelor of Science in Material Physics — *Aug 2013 – May 2017*