Amir Sohel

Staff Technical Program Manager | Agentic AI Platforms • Cloud Infrastructure • 0 to 1

Toronto, Canada • aot.amirsohel@gmail.com • linkedin.com/in/asohel • github.com/amir-sohel-builds • amir-sohel-builds.github.io
Canadian citizen • TN eligible • Open to U.S. relocation

SUMMARY

Staff Technical Program Manager and former software engineering lead with 12 years delivering AI and cloud platforms, 7 of them in technical program management. Led an end-to-end agentic AI platform from 0 to 1 as sole TPM and scaled it to 79+ production agents across HBO Max, Discovery+, TNT Sports, and CNN, where technical and non-technical teams deploy and operate their own agents and apps. Hands-on with Claude Code, the Claude API, and Amazon Bedrock. Supports streaming products serving 130M+ subscribers.

CORE EXPERTISE

AI platforms: Agentic AI • Claude API • Amazon Bedrock • AgentCore • MCP and tool gateways • RAG and semantic layers • IAM and OAuth • Prompt lifecycle • Evaluation, observability, cost tracking
Program leadership: 0 to 1 strategy • Roadmap tradeoffs • Cross-org delivery • Operating models • KPIs/OKRs • Decision rights • Risk and escalation
Cloud and engineering: AWS • Kubernetes/EKS • Terraform • Argo CD/GitOps • CI/CD • Kafka • Java • Python • SQL • GraphQL • REST APIs

EXPERIENCE

Staff Technical Program Manager | Warner Bros. Discovery

Toronto | Feb 2025 to Present

Sole TPM for the Platform Engineering organization behind HBO Max and WBD streaming services globally.

- **79+ agents live in production.** Led the agentic AI platform from 0 to 1, from concept to launch, so technical and non-technical builders can get agents and apps deployed and operated. Drove adoption across the engineering and product teams behind HBO Max, Discovery+, TNT Sports, and CNN, and turned engineer discovery into the roadmap, operating model, and KPIs.
- Delivered the full platform: Bedrock inference profiles and a model harness, AgentCore runtime and deployment, IAM and 3LO, a governed gateway for MCP and tool connectivity with domain-level security, RAG over vector databases and knowledge bases with a semantic layer for querying data, and observability with cost tracking. Releases are eval-gated, with prompt versioning and rollback.
- Built the Platform Engineering org's operating system: a planning portal behind SSO, a marketplace of 150+ agent skills (40+ authored personally) with a low-code builder for PMs, 20 Claude Code agents, and eight live dashboards.
- Authored a source-first guardrail plugin any team can install: agents check configured sources first and refuse rather than invent answers. Also built a token-routing plugin that cut inference cost about 66%.
- **Cut cloud spend 43% in a year** by leading decommissioning across 40+ AWS accounts, sustaining 99.995% uptime through peak streaming events.
- Raised SLO compliance from 67% to 84% and cut support tickets 28% after analyzing 17,000+ tickets. Self-service deployment cut code-to-production from three days to under four hours, with one-click rollbacks.

Senior Technical Program Manager | Xactly Corp

Toronto | Jul 2022 to Feb 2025

Platform transformation for an enterprise SaaS company with 500 customers and six product teams.

- Led an 18-month identity transformation across six product teams, moving 500 enterprise customers to Okta SSO and reducing login latency 60%.
- Built the roadmap for a phased AWS EKS migration, cutting infrastructure costs 35% and release cycles from four days to same-day with Argo CD GitOps.
- Stood up a cross-functional AI and ML steering committee across engineering, product, and data science, shipping churn prediction and revenue forecasting.
- Migrated business-critical databases to Azure with 99.9% uptime and zero data loss, and production pods to Oracle Cloud Infrastructure, cutting annual infrastructure cost.

Lead Software Engineer, then Technical Program Manager | Cognizant

Toronto | Apr 2014 to Jul 2022

Backend lead (Apr 2014 to Nov 2019), then TPM (Dec 2019 to Jul 2022) for payments, point-of-sale, and e-commerce at a Fortune 100 retailer.

- Led delivery of a large e-commerce program and scaled a payment platform from a 10-store pilot to 5,000 locations, at peak Black Friday loads of 50,000 transactions per minute and 99.95% uptime, coordinating store systems, payment processors, and PCI compliance.
- Architected a GraphQL layer that cut API response time from 800 ms to 200 ms and payload size 70%. Built Java and Spring Boot services for real-time inventory and pricing, sustaining 10x normal traffic during promotions.
- Incident commander for Sev1 outages: led triage, coordinated resolution, and wrote RCAs. Mentored 3 engineers, 2 promoted within 18 months.

EDUCATION

Bachelor of Technology, Electronics and Instrumentation Engineering | West Bengal University of Technology