

Bharath Parameshwara

Seattle, WA | bharathp024@gmail.com | +1 937 789 2529

linkedin.com/in/bharath-parameshwara | Open to SF / NYC and travel

Senior Software Engineer | Applied AI and Agent Systems

10+ years shipping production systems. Currently building LLM-assisted advertising experiences at Amazon Ads; previously led AI safety engineering and enterprise delivery at AWS. Solo-built Atlas, serving 15,600+ chat sessions in 28 days. Owns work from customer needs and architecture through implementation, evaluation, and launch.

Experience

Amazon | Senior Software Development Engineer

Amazon Ads | Mar 2026–present

- Building premium brand experiences on product detail pages, using LLMs to generate and refine product titles, qualifiers, and other advertising content.
- Designing agent-assisted workflows for campaign creation and ad experience design.

Amazon Web Services | Senior Software Development Engineer

Nov 2023–Mar 2026

- Architected multi-model LLM safety for multi-modal production workflows; partnered with the AppSec red team to harden the service against common LLM attack classes. Handled hundreds of requests per second in production; load-tested at thousands of requests per second.
- Led enterprise PII detection and redaction: 8 engineers, 5 teams, 18 weeks, and 4 modalities. Worked with 4+ enterprise customers and solutions architects from solution design through beta launch.
- Cut safety latency 60–70% through concurrent calls, direct payload support, and model/provider migration; reduced safety API cost and inference fleet size by 50%. Maintained 99.9% availability.
- Operated a multi-region GPU inference fleet with hundreds of instances per region; built tail-latency, memory, and thread-utilization observability. Authored the safety roadmap across cost, latency, compliance, and technical debt.
- Led region automation that cut commercial launches from 1–2 months to under 2 weeks across 5+ new regions, covering infrastructure, canaries, alarms, and compliance.
- Helped grow the team from 5 to 17 engineers through 25+ interviews; mentored 4 SDEs and 1 intern and established on-call and incident-response procedures.

Amazon Web Services | Software Development Engineer II

Mar 2022–Nov 2023

- Built shared ML inference libraries; introduced continuous deployment that reduced code-to-production time by 50% and improved job reliability.

Earlier Experience

Amazon Advertising | Software Development Engineer

Feb 2021–Mar 2022

- Built and launched highly available APIs for advertising workflows; led project design and deployment automation.

Accenture | Application Development Senior Analyst

Sep 2014–Jan 2019

- Delivered Go and Kubernetes microservices 30% ahead of schedule; led a C#-to-Java migration and a six-developer team.

Selected Projects

Atlas | AI travel planner

Mar 2026–present

- Solo-built the full stack and launched atlas.meetedith.com: 15,600+ chat sessions, 4,200+ itineraries, and 5,400+ refinements over 28 days (Sep–Oct 2026).
- Built chat-driven planning, scoped edits, saved trips, and grounded travel answers with Next.js, TypeScript, and Postgres; shipped on Vercel, Neon, and AWS.

Edith | Personal AI assistant

Mar 2026–present | Private beta

- Built WhatsApp-native orchestration with 18+ tools, fast and deep workflows, and maintained vector memory. Routing benchmark: 18/21 cases passed at 2.18s median latency.
- Built provider-agnostic routing, fallback, per-call cost attribution, and quality/latency benchmarks; self-hosted vLLM inference and local embeddings for cost-sensitive paths.

Cortex | Knowledge infrastructure

Mar 2026–present

- Built a Python/FastAPI and Postgres retrieval service with 300K+ items across 6 domains; gated LLM entity reconciliation with golden evaluations and hard negatives.

Project S | Market research platform

Aug 2026–present | Personal use

- Built agents that extract, embed, rank, and validate trade ideas from video, community, and web sources. Durable Postgres jobs, market-aware scheduling, and duplicate suppression replace repetitive manual research.

Technical Skills

Applied AI: Agents, tool use, RAG, vector memory, model routing, fallback, safety evaluations; Claude, Gemini, GPT, open models, Strands SDK.

Product and data: TypeScript, Python, Next.js, React, FastAPI, Postgres, Prisma, WhatsApp Cloud API; Pinecone, Qdrant, Vectorize.

Infrastructure: AWS, Triton, vLLM, Docker, Kubernetes, Vercel, Cloudflare; Java, Go, SQL.

Education

MS, Computer Science | University of Dayton | GPA 3.78 | Jan 2019–Dec 2020