Shreshth Kharbanda

shreshthkharbanda1@gmail.com | (404) 510-6272 | kharbandas.com | linkedin.com/in/skharbanda/

SKILLS

Core Languages: Java, Python, C++
ML & Inference: PyTorch, TensorFlow, vLLM, FlashAttention, SageMaker, Triton, Transformers
Cloud & Infrastructure: AWS (ECS, Fargate, API Gateway, Lambda, SageMaker, Bedrock, CDK), Docker, Terraform

PROFESSIONAL EXPERIENCE

Software Engineer 2, Amazon Web Services *June 2023 - Present*

- Launched a new API introducing guardrails for agents with a tool adherence policy. Interacted with **50+ customers** to scope and drive integrations enabling safe production usage of agentic AI.
- Launched Bedrock Guardrails Standard Tier in partnership with research and product driving **30%+ accuracy gains**, **25% higher GPU utilization** via cross-region inference, expansion to **60+ languages**
- Defined and owned a model-agnostic integration framework reducing model onboarding from **17 days → 1 day** and enabling **40+ model launches** at re:Invent
- Re-architected stream buffering and evaluation improving **time-to-first-token by 67%** with equivalent accuracy
- Architected tier-aware admission control to mitigate noisy-neighbor effects during burst traffic, **improving service availability by 15%** under load shedding leveraging Redis + Lua
- Designed and implemented multi-endpoint slot-pooling architecture and dynamic batching for inference improving **throughput by 3x** and **reducing p95 latency by 50%**

Software Engineer Intern, Tesla *January 2023 – June 2023*

- Implemented features for Tesla's mobile app **serving 60k+ weekly users**, partnering with product and design
- Led development and launch of a customer-facing installer discovery service supporting 20k+ weekly users
- Implemented Redis-based caching for the backend-for-frontend, **reducing page load latency by 65%**

Research Assistant, University of Washington *April 2022 - June 2023*

- Internationalized a research-backed educational social media platform with **500,000+ users** enabling scalable, privacy-compliant data collection
- Built distributed data pipelines and supervised learning models (regression + classification) to identify engagement bottlenecks and predict high-friction modules, **reducing manual analysis by 1,000+ hours**

Software Engineer Intern, Chewy *June 2022 – August 2022*

- Applied incremental static regeneration for a micro-frontend to **reduce load speeds by 300%** for 20M+ users

Software Engineer Intern, CodeLabs *June 2020 – August 2020*

- Developed an Android misinformation-detection app using Java and REST APIs, tested with 30+ beta users
- Designed and trained a TensorFlow NLP model for bias detection achieving 77% accuracy on labeled news article data

Founder, JoDi Services *April 2020 – June 2022*

- **Drove 25+ digital transformation projects** using React.js, React Native, Java, Node.js, Python, AWS, GCP

PROJECTS

Atma: Personal Assistant on iMessage (atma.bot)

- Built a personal agent that acts as an executive assistant interfacing through iMessge using Java SpringBoot + MCP

Contour: Autonomous Engineering Agent

- Built a long-running AI agent that operates continuously across development and operational workflows, autonomously planning work, implementing changes, reviewing code, and investigating using persistent memory and structured code context
- Automates **25+** hours/week and reduces on-call triage from 30 → 7 minutes with **85%** root cause accuracy
- Produces PRs that are **40% mergeable on first revision** and converge within ~3 iterations on average

Curiosity: Multi-Threaded Search Engine

- Built a search engine in C, C++ with dynamic indexing processing 10,000+ documents
- Optimized tool using multithreading, big-endian handling, and architecture-neutral data marshalling
- Enabled secure and concurrent request handling using Boost, STL to reduce server response time by 30%

EDUCATION

B.S. Computer Science, University of Washington – Seattle

- Coursework: Distributed Systems, Hardware Systems, Deep Learning, AI, ML, Data Structures & Algorithms, Software Design, Databases
- Organizations/Clubs: Algorithmic Trading Club, DubHacks, Computing Community