Vedant Bhat

Seattle, WA • vbhat3443@gmail.com • 678.702.6107 • U.S. Citizen
https://www.vedantbhat.co • https://www.linkedin.com/in/vedant-bhat

SUMMARY

SDE II and Georgia Tech MSCS graduate with experience designing and building **scalable distributed systems and low-latency AI/ML pipelines**. Proven expertise in delivering software that supports **real-time AI inference across large-scale data systems.**

EXPERIENCE

AMAZON **Seattle, Washington**

Software Development Engineer II, AWS Vulnerability Management Oct 2025 – Present

- Developing a real-time rules engine **processing 3B+ Host/CVE evaluations/day** (~35K TPS), improving vulnerability detection speed and reducing exposure to critical security threats across AWS systems.
- **Owned end-to-end system design**, defining service boundaries, data contracts, traffic management, and resilience strategies to support reliable, low-latency operation at sustained high throughput.
- **Leveraged task-specific LLMs** to recalibrate CVSS scores based on contextual risk, increasing security posture by escalating critical threats while reducing builder fatigue by filtering out low-risk false positives.

Software Development Engineer I, Amazon Personalization Jan 2024 – Sep 2025

- Architected and deployed an LLM-powered content moderation and curation pipeline for Amazon Homepage, Search, and Detail Page, increasing throughput from **144K/year (manual) to 13M/year** and delivering **~$700K in annual cost savings.**
- Served as a core engineer on Amazon's Content Recommendations Service, implementing request caching with Memcached and optimizing service logic to cut P99 latency by ~50% (79ms → 41ms) under Prime Day-scale traffic (~10K TPS).
- Led the training and deployment of two ML classifiers using XGBoost: one predicting content engagement and one predicting human curator approval, improving content ranking and moderation decisions in Shop-by-Interest feeds.
- Built a semantic content search system **indexing 1.2M+ photos and videos** using large-scale embeddings and vector search to surface high-quality content recommendations for complex customer search queries.

Software Development Engineer Intern, Generative Media May 2022 – Aug 2022

- Engineered an automated service to streamline the creation of a Visual Review (VR), an innovative advertising format within Amazon's retail catalog, by leveraging AWS services, including CDK, API Gateway, Lambda, and S3.
- Designed and delivered an automated workflow that resulted in the dynamic generation of SVG images by programmatically combining a background SVG and a popular product review, streamlining a previously manual process.

IBM **RTP, North Carolina**

Software Engineer in Test Intern, DataPower May 2021 – Dec 2021

- Automated regression and integration testing for IBM's DataPower Gateway using Java, JUnit, and REST Assured, integrating with Jenkins CI/CD pipelines to validate APIs and configurations while cutting manual QA effort by 120 hours annually.
- Created scalable test frameworks that enhanced coverage of core features, supporting CI/CD and reducing QA bottlenecks.

FISERV **Alpharetta, Georgia**

Software Engineer Intern, Debit Routing Systems May 2020 – Aug 2020

- Built a JavaFX simulation platform enabling engineers to design debit routes and test real-time payment processor integrations.

SKILLS

- **Programming Languages:** C++, Java, Python, SQL, TypeScript, Rust, C, Scala
- **Cloud Infrastructure:** AWS (ECS, Lambda, S3, DynamoDB, SageMaker, Step Functions, CloudWatch, CDK, IAM, API Gateway, Glue), Kubernetes (EKS), Docker (ECR)
- **AI & Data:** PyTorch, Pandas, RAG, Vector Databases, Semantic Search, LLM Integration, Model Training/Evaluation
- **System Design & Architecture:** Distributed Systems, Event-Driven Architecture, Stream Processing (Kafka/SQS), Microservices (REST, gRPC, GraphQL), Fault Tolerance, Caching Strategies (Redis/Memcached)
- **Operational Excellence:** Scalable Request Handling, Observability and Metrics, Testing Frameworks, Failure Recovery
- **Graduate Coursework:** Operating Systems, High Performance Computing Architecture, High Performance Computing Algorithms, Artificial Intelligence, NLP, Algorithms, Computer Graphics, Scientific ML, High-Dimensional Data Analytics

EDUCATION

GEORGIA INSTITUTE OF TECHNOLOGY, College of Computing **Atlanta, Georgia**

Master of Science in Computer Science Dec 2024

- Specialization: Machine Learning and Computing Systems, Scholarships: Terrill Graduate Fellowship GPA: 4.0

Bachelor of Science in Computer Science May 2023

- Specialization: Artificial Intelligence and Information Internetworks, Scholarships: Zell Miller GPA: 3.8