# Sahil Hamal

Full Stack / Senior SWE / Gen AI | www.sahilhamal.dev | Reston, VA | sahilhamal@vt.edu

## Skills

**AI / GenAI,** LLM integration (OpenAI API, LLaMA, Claude), RAG pipelines, semantic search, vector embeddings, prompt engineering, LLM evaluation, Hugging Face Transformers, PyTorch, agentic workflows (ReAct, LangChain)

**Languages,** Python, Java, JavaScript, TypeScript, SQL, C#

**Backend / APIs,** Spring Boot, Node.js, REST, GraphQL, microservices

**Frontend,** React, Vue.js, React Native, Angular, HTML5, CSS

**Data/Search,** ElasticSearch, Redis, vector stores, MongoDB, PostgreSQL, MySQL

**Cloud,** AWS (Lambda, S3, API Gateway, Bedrock-adjacent workflows), Docker, Kubernetes, CI/CD, GitHub Actions

## Work Experience

**Software Engineer II,** American Express — 06/2023 – Present | Reston, VA

- **Designed and shipped an LLM-powered transaction categorization system** (OpenAI API + embedding-based retrieval for few-shot context) that replaced a legacy rules-based pipeline, improving classification accuracy from 78% → 94% measured on a held-out production dataset of ~50K labeled transactions.
- **Built a retrieval-augmented generation (RAG) pipeline** over internal transaction metadata using LLaMA embeddings + ElasticSearch as the vector/keyword hybrid store, served via Spring Boot; reduced p95 query latency ~40% through Redis caching and re-ranking optimization.
- **Established LLM evaluation framework** with precision/recall tracking, prompt regression tests, and A/B comparisons across model versions; used telemetry to iteratively tune prompts and retrieval parameters based on measurable impact.
- **Optimized LLM inference tradeoffs** (model choice, token budget, batching, caching) to balance latency vs. accuracy, cutting per-request cost while maintaining classification precision above SLA.
- **Led design reviews and code reviews** across a 3-engineer team, driving production rollout decisions, architecture proposals, and safe deployment patterns for AI features in a regulated financial environment.
- **Partnered cross-functionally** with Product and Analytics to translate business requirements into AI features;resolved ingestion/search inconsistencies with idempotent pipelines and validation checks to ensure integrity in financial analytics workflows.

**Full Stack Developer Intern,** Zoom Video Communications — 05/2022 – 08/2022 | San Jose, CA

- **Built** a full-stack video archive for Zoom Events with tag-based ElasticSearch retrieval and real-time indexing across thousands of event recordings — an early retrieval-system foundation for search-over-content work later extended with LLMs.
- **Architected** hybrid AWS S3 + MongoDB storage with CDN delivery and an ElasticSearch metadata pipeline enabling sub-second full-text search, tag filtering, and playback across archived recordings.

**Software Engineer,** Johns Hopkins University — 10/2019 – 06/2021 | Baltimore, MD

- **Application developer for AstroPath** — an NIH-funded cancer imaging platform published in *Science* (June 2021), covered by *The Economist*, and recipient of the 2021 Falling Walls Life Sciences Award (https://astropath.org/team ).
- **Built the full-stack cell-view annotation and cancer detection interface** used by JHU immunologists under NIH data governance and IRB compliance; contributions supported securing 5+ years of continued research funding.

**Software Engineer,** Deutsche Bank — 06/2018 – 10/2019 | Cary, NC

- **Developed** a real-time dependency tracking system for financial communication platforms with automated notifications, preventing compatibility failures across enterprise systems during scheduled release windows.
- **Designed** an automated Ansible-based deployment pipeline for financial applications into secured, compliance-governed banking environments with full audit trail and rollback.

**Software Developer Intern,** PricewaterhouseCoopers (PwC) — 06/2017 – 07/2017 | Dallas, TX

- **Built a financial account management application** for PwC Tax & Technology with report generation and redundancy reduction, delivered within a Big Four professional services environment.

## Education

**Masters in Computer Science,** Virginia Tech — 08/2021 – 05/2023 | Blacksburg, VA

Concentration: Visual Analytics & Explainable AI | Advisor: Dr. Chris North

*Thesis: Interpreting Dimension Reductions Through Gradient Visualization* | Award: **Torgersen Outstanding Graduate Research Award**

**Bachelors in Computer Science,** Troy University — 08/2014 – 05/2018 | Troy, Al

President of Computer Science Club | University Ambassador