Kashyap M Kale

San Francisco, CA

+1 571-461-9423 | kashyapk@vt.edu | linkedin.com/in/kashyapmkale

https:/github.com/kashyapkale | coding-kashyap.vercel.app

Summary

Experienced in backend software engineering with comprehensive expertise in Java and cloud computing technologies. Notable accomplishments include optimizing system performance for Tray.com and engineering data ingestion pipelines at Amazon, enhancing efficiency and accuracy. Eager to leverage skills in system integration and cloud platforms to drive successful projects and innovation.

Education

Virginia Tech **Jul 2024 - Present**

Master of Engineering, Computer Science (GPA: 3.9/4.0) *Alexandria*

- **Coursework:** Multiprocessor Programming, Cloud Computing, Machine Learning, AI Tools, NLP

Maharashtra Institute of Technology **Jul 2017 - Jul 2021**

Computer Science and Engineering (GPA: 8.76) *Pune*

- **Coursework:** Operating Systems, DBMS, Microprocessors, OOP, DSA

Work Experience

Virginia Tech | *Student Software Engineer* **Aug 2025**

- Built a production-ready retrieval-augmented generation pipeline using S3 Vectors for semantic search across canvas.
- Implements advanced two-phase retrieval with hybrid scoring (vector similarity + keyword overlap), smart note routing, and context-aware answer generation achieving 85%+ answer accuracy on course-specific queries.
- Designed and implemented a sophisticated multi-step agent workflow where the LLM autonomously decides when to call tools. Features role-based tool access, multi-turn orchestration, and graceful handling of ambiguous queries with contextual clarification requests.
- Developed a full-stack solution with Spring Boot REST API (Java), Flask microservice (Python), and AWS Lambda vectorization workers. Implements JWT-based authentication, Canvas LMS integration for real-time grade sync, PostgreSQL persistence, and AWS S3 for document management with automated chunking and embedding pipelines processing 500+ pages of course material.
- Engineered comprehensive security boundaries and guardrails where students access only their own grades, while instructors use multi-step verification to prevent unauthorized data access.
- Successfully rolled out to 5 students in pilot phase with positive engagement metrics. System designed for horizontal scaling with stateless microservices, async vectorization workers, and efficient vector indexing supporting 10K+ queries/day. Scheduled for department-wide deployment next semester serving 200+ students across multiple courses with per-course index isolation and instructor-specific administrative tools.

Tray.com | *Software Engineer* **Apr 2023 - Aug 2024**

- Optimized Java, Spring Boot, and RESTful APIs to enhance system performance by 15% and elevate customer satisfaction by 10%.
- Led the integration of the 7Shifts platform with Tray's Spring Boot microservices, REST APIs, and POS system, diminishing manual workload by 25% for 50+ clients.
- Reduced cloud bandwidth usage by 30% and minimized data inconsistencies through efficient binary serialization and in-memory batching strategies.
- Deployed services via Docker on AWS EC2 and established production monitoring with Prometheus metrics and ELK-based logging for observability and root-cause debugging.

Amazon | *Software Development Engineer* **Mar 2022 - Apr 2023**

- Developed and maintained backend infrastructure for a multimedia data labeling platform within the Alexa Data Services team, processing over 200,000 unlabelled data items monthly using AWS Lambda, Amazon SQS, and Amazon SNS.
- **Designed and implemented** the callback orchestration layer for SageMaker Ground Truth and Mechanical Turk, enabling external systems to securely resume paused SageMaker Pipelines via token-based APIs.
- Implemented callback orchestration and token validation services for SageMaker Ground Truth and MTurk, enabling secure, event-driven human-in-the-loop workflow integration within SageMaker Pipelines.
- **Enhanced scalability and fault tolerance** of multi-service ML labeling pipelines, supporting **millions of callback events monthly** and ensuring consistent pipeline resumption across distributed AWS regions.
- Designed and implemented a quality computation system comparing annotations from three different data associates per item, reducing labeling errors by 15% and improving dataset accuracy for machine learning models.

Projects and Publications

AI-Powered Restaurant Kiosks with LLM and RAG Integration **Oct 2024 - Jan 2025**
https://github.com/Kiosk-ByteCrew

Virginia Tech

- Developed and implemented a AI solution for self-ordering kiosks to enhance user interaction by integrating large language models (LLMs) with natural language capabilities.
- Designed and deployed a retrieval-augmented generation (RAG) system using OpenAI APIs, enabling kiosks to deliver personalized, context-aware, and accurate responses to customer queries.
- Integrated Whisper for speech-to-text processing, allowing users to interact with the kiosks using natural spoken language, further improving accessibility and ease of use.
- Enhanced the system by replacing OpenAI APIs with **llama 3.2** and integrating **FAISS for semantic search** of menus and ingredients, while also incorporating **BM25 for keyword search** to improve retrieval accuracy.

Building Applications using OpenAI APIs(Publication) | DOI : 10.1007/978-981-97-8460-8_3 **Apr 2024 - Jul 2024**
Studies in Computational Intelligence, Springer

- Detailed API Features: Explained core parameters like temperature, max tokens, and frequency penalty, with best practices for tuning model outputs.
- Movie Recommendation System: Developed an AI-based recommendation system that provides tailored suggestions through natural conversation.
- Customizable Chatbot: Built a versatile chatbot for various use cases, supporting dynamic responses and task-specific functions.

Skills

- **Programming Languages:** C++, Java, Javascript, TypeScript, Object Oriented Programming
- **Cloud Platforms:** AWS, Azure, GCP
- **API & Integration:** Spring Boot, gRPC, GraphQL, LLMs, AWS Bedrock
- **Development Tools:** Intellij Idea, Cursor, Claude Code

Achievements

- **DSA:**2024 Solved over 600 data structure and algorithm problems in C++ on various sites.
- **LTI ICC:**2022 1st place in the LTI Infinity Coding Challenge (800+ participants).
- **Tech Hunt:**2019 1st position in Tech Hunt coding competition organized by Texyphyr Pune.
- **Maths:**2017 Center topper in Mathematics (95/100).