

# ALISHA VINOD

(347) 847-6330 | av3311@columbia.edu | www.linkedin.com/in/alishavinod | www.github.com/alishavinod

## EDUCATION

### Columbia University

New York, NY

#### Masters of Science in Computer Science (GPA - 3.87/4)

Expected Dec 2026

Operating Systems, Programming Languages and Translators, Natural Language Processing, High Performance Machine Learning

### National Institute of Technology, Warangal

#### Bachelor of Technology in Electronics and Communication Engineering (CGPA - 8.63/10)

Telangana, IN

Data Structures and Algorithms, Computer Architectures, Problem Solving and Programming, Data Networks.

May 2022

## TECHNICAL SKILLS

**Languages & Frameworks:** C, C++, TypeScript, JavaScript, C#, Python, SQL, React, REST APIs, .NET MVC, Kubernetes, Flask, Java

**AI/ML & Infrastructure:** RAG, LLM fine-tuning, LangChain, PyTorch, TensorFlow, CUDA, GCP, AWS, Docker, Azure, PostgreSQL

**Other:** Git, GitHub, OOP, Compiler Design, Linux, Data Structures and Algorithms, Distributed Systems, System Design, Claude Code

## WORK EXPERIENCE

### Aptean

Alpharetta, GA

#### Software Engineer Intern

May 2026 - Present

- Instrumented product analytics with PostHog across 7 key user flows, surfacing engagement insights that drove product and AI roadmap decisions; fine-tuned the open-source PostHog AI wizard model, improving code coverage of capture logs by 75%.
- Architected an Azure Kubernetes Service (AKS) platform running self-hosted PostHog as 20+ microservices including ClickHouse, Redis, Kafka, MiniIO, PostgreSQL and other stateless services fronted by Cloudflare supporting 10k users on App Central.
- Engineered AI-powered features for Aptean Intelligence Studio (AIS), including a voice agent developed on the GPT-Realtime 2 model, improving conversational reliability and response quality for end users.

### Mastercard

Pune, IN

#### Software Engineer II

Jul 2022 - Jul 2025

- Delivered scalable UI components in React-Redux and TypeScript integrated with REST APIs in .NET C# to enhance usability of Test & Learn, a large-scale analytics platform serving 300+ clients and driving \$351M in revenue (+23% YoY).
- Owned end-to-end system design and delivery of new user-facing features independently, from design through CI/CD deployment with Jest and NUnit test coverage including an n-gram ranked search that reduced time on platform by 17%.
- Debugged and resolved 40+ production issues through RCA, Splunk monitoring and collaborating cross-functionally with clients.
- Designed a Retrieval-Augmented Generation (RAG) based conversational chatbot leveraging LangChain and OpenAI GPT-3.5 Turbo to answer user queries through confluence documents reducing manual support effort by 65%.

## PROJECTS

### Retail Customer Service Agent

Mar 2026 - Apr 2026

- Implemented a customer-service agent with four tools (order lookup, cancellation, modification, payment) across two frameworks, LangGraph and the OpenAI Agents SDK, to compare orchestration approaches on the same task.
- Evaluated both implementations on a 10-task benchmark measuring correctness and tool-calling accuracy; extended the system with a TravelPilot agent reaching a 0.94 average evaluation score, alongside a 0.4 pass@1 on the core task set.

### MacroCart - Personalized Nutrition Planning Agent

Nov 2025 - Dec 2025

- Developed an LLM-based personalized nutrition assistant that generates meal plans and recipes with macros from user profiles.
- Designed a RAG pipeline over ingredient, macro, and pricing datasets, reducing variance by 35% versus a prompt-only baseline.
- Benchmarked Llama 3 vs. GPT-4.1-mini and integrated the Instacart API to generate shoppable grocery lists from recommended meals, achieving 30% lower macro-target deviation.

### HPML - TSFM Agent Inference Optimization (AssetOpsBench-IBM)

Mar 2026 - Apr 2026

- Profiled and optimized the inference pipeline of an IBM TinyTimeMixer (TSFM) agent within a multi-agent MCP benchmark, reducing end-to-end workflow latency 68% (43,021ms to 13,798ms).
- Eliminated cold-start overhead via model caching, cutting model load time 9x (233ms to 26ms); applied torch.compile kernel fusion to collapse 19,032 per-call CUDA kernel launches, cutting forward-pass latency 69%.
- Benchmarked GPU placement, TF32, and bf16 mixed-precision configurations across 9 sensor workloads using latency profiling, PyTorch Profiler, TensorFlow for GPU utilization charts and Weights & Biases, achieving a 3.3x total speedup over CPU baseline.

### Systems & Compiler Projects

Sep 2025 - Nov 2025

- Built an end-to-end compiler (lexical analysis, AST-based parsing, semantic analysis, machine code generation) validating correctness across 30+ test programs with explicit control flow and memory management.
- Designed a custom Oven scheduler in the Linux kernel to optimize task prioritization, concurrent processing, load balancing, and CPU fairness, achieving 20–25% lower average task completion time than default fair scheduler.