Vineet Agarwal

240-353-9811 — vineetagarwal540@gmail.com — vineetagarwal54 — Github — Portfolio

Summary — Software engineering grad student with 2+ years production experience, focused on multi-agent AI systems and LLM inference optimization, from agent orchestration (LangChain, LangGraph, Bedrock) down to CUDA kernel fusion and 4 bit quantization on GPU workloads.

Skills

LLM Inference & Optimization vLLM, SGLang, llama.cpp, CUDA, quantization, speculative decoding, KV cache

Applied AI LangChain, LangGraph, AutoGen, RAG pipelines, OpenAI APIs, Amazon Bedrock, Hugging Face

Databases PostgreSQL, MongoDB, Redis, FAISS, Pinecone

Cloud & DevOps AWS (Lambda, EKS, Docker, Kubernetes, CloudFormation, GitHub Actions CI/CD

Languages Python, JavaScript, TypeScript, C++, SQL, Go

Frontend React, Next.js, React Native, Redux Toolkit

Backend FastAPI, Node.js, Express, Django, REST APIs, WebSocket, WebRTC

Core Engineering Data Structures & Algorithms (C++), OOP, System Design, Microservices, Agile/Scrum

Experience

ServBeyond Solutions **Jun 2026 – Present**
Enterprise AI & Platform Intern

- Built and shipped a RAG assistant over internal documentation using LangChain and OpenAI APIs, cutting 25 hours of manual lookup per week and reaching 95% answer accuracy across 100 users.
- Designed agentic GenAI workflows across Salesforce, ServiceNow, and AWS using Amazon Bedrock for agent orchestration and REST plus SOAP integration into a normalized data layer.

Runara.ai **March 2026 – May 2026**
Software Engineer Intern

- Built a reproducible benchmarking pipeline to compare vLLM, SGLang, and TensorRT-LLM serving Qwen3-Coder-480B across 1K to 128K context on RTX Pro 6000 Blackwell GPUs, profiling TTFT, throughput, KV cache, and runtime stability.
- Implemented fused RMSNorm and Linear CUDA kernels for the Qwen3 attention projection, validated to within 7.5e-6 of the reference at fp32, removing a redundant normalization pass per layer.

Xelpmoc Design and Tech Limited **Nov 2022 – Apr 2024**
Software Engineer

- Reduced API response time from 75s to under 10s and increased throughput 4x by refactoring SQL joins, indexing high-traffic tables, and adding a Redis caching layer.
- Architected RESTful backend services for a tourism platform serving 10K+ daily bookings; cut production defects 40% and accelerated feature delivery 25%.

Svipes **Jul 2024 – Dec 2024**
Software Engineer

- Shipped 50+ React Native screens with Redux Toolkit and cut video playback stutter 70% via async loading and client side caching.

IIIT Hyderabad **Jun 2022 – Aug 2022**
Software Intern

- Automated multilingual data extraction across 5+ regional languages with Python and Selenium; lifted NLP classification accuracy from 78% to 93%.

Education

University of Maryland, College Park **Jan 2025 – Dec 2026 (Expected)**
Master of Engineering, Computer Software Engineering — Minor in Cloud Engineering *GPA: 3.7*

Projects

RepoResearchAI : Multi-Agent Code Analysis Platform (Github)

- Built a multi-agent AI system with AutoGen, LangChain, and FAISS vector search; specialized agents (SDE, PM, QA) generate architecture insights, API docs, and schema analysis using GPT-4 with RAG over codebase embeddings.

Care Bridge AI (Github)

- Built a healthcare assistant at Bitcamp using FastAPI, PostgreSQL, React, and the Gemini API, delivering conversational patient support with low-latency responses.