Abi Kakolla

Los Angeles, CA | 323-641-2494 | kakolla@usc.edu | kakolla.com | kakolla | kakolla

Education

University of Southern California — Los Angeles, CA
Bachelor of Science in Computer Science — *Aug 2023 – May 2027*
Relevant Coursework: Data Structures & Algorithms, Theory of Computation, Software Engineering, Computer Systems
Leadership & Activities: Viterbi Conversations in Ethics, CURVE Researcher, Open Alpha Games

Experience

Software Engineering Intern - ML & Infrastructure — May 2025 – Present
StitcherAI — Redmond, WA

- Architected LLM inference simulation models with Meta's Llama 4 Maverick to predict enterprise-scale FinOps workloads for Walmart and Indeed
- Deployed a Redis-integrated LlamaIndex reflection workflow handling Polars LazyFrames to cache agent memory, reducing query latency for large financial datasets ($500M+ in cloud spending)
- Developed a graph-state ETL operator in TypeScript connected to a FastAPI backend to track client data transformations
- Optimized graph scalability with BFS traversal + hash map, efficiently handling 10K+ requests for enterprise workloads

Software Engineering Intern — Feb 2025 – May 2025
FoundrySix (partnered with Meta) — Los Angeles, CA

- Built Gemini LLM-driven Augmented Reality agents for a Meta-funded project on the Quest platform, facilitating natural language interaction within immersive environments
- Implemented asynchronous C# server-side logic to dynamically load .FBX assets and sync data with DynamoDB in real time, reducing load times by 40% for 200+ concurrent users
- Designed game environments and optimized lighting performance on the Meta Quest by precomputing baked lighting

Machine Learning Researcher — Aug 2023 – Present
USC Center for Neural Engineering — Los Angeles, CA

- Designed a multi-agent GraphRAG pipeline (Python, LangChain, Neo4j) to improve LLM reasoning and cut hallucination 10× for 120+ computational neuroscience papers with a chain-of-thought graph-based architecture
- Optimized cortical axon generation accuracy and cut down compute costs by 30% through code profiling and software-level optimization, enabling faster neural pathway modeling for research scientists
- Decreased neural network generation runtime by 75% by implementing K-means clustering with Scikit-learn and NumPy

Projects

C Compiler (AST, Lexers/Parsers) | *C++* — Link

- Implemented a lexer and recursive-descent parser to generate Abstract Syntax Trees (ASTs), supporting arithmetic operations, variable assignments, and function definitions
- Constructed a code generation backend targeting ARM assembly, with a C++ testing framework to validate correctness, allowing compilation and execution of C programs

Clean Sweep – Harvard Hackathon Winner | *Python, OpenCV, Terraform, Databricks* — Link

- Engineered a smart route prediction platform (React, NodeJS, Databricks), winning Hack Harvard 2024 (150 teams) for best use of Terraform
- Wrote an OpenCV contour-detection algorithm and Random Forest Classifier deployed with Kubernetes + Terraform, enabling real-time trash level detection and optimized route predictions

Published Abstracts

J.-M C. Bouteiller, **A. Kakolla** et al. "**From Data Integration to Discovery: A Transparent, AI-Driven Framework for Multidisciplinary Brain Research.**" *Society for Neuroscience 2025*, San Diego, CA.

Skills

Languages: C++, Python, C#, Java, Lua, JavaScript/TypeScript, SQL, Cypher
Frameworks: FastAPI, NumPy, Scikit-learn, LlamaIndex, LangChain, Pydantic, React, Next.js, Node.js
Technologies: Polars, Redis, Neo4j, Docker, Kubernetes, Terraform, Databricks, Grafana