Soham Pati

678-599-7537 | sohampati2005@gmail.com | linkedin.com/in/soham--pati | github.com/sohampati | U.S. Citizen

Education

Georgia Institute of Technology — **Aug 2023 – May 2027**
Bachelors/Masters in Computer Science - GPA: 3.7/4.0 — Atlanta, GA
Concentration: Intelligence (Machine Learning & AI), Info-Networks
Relevant Coursework: Data Structures & Algorithms, Advanced Algorithms, Object-Oriented Programming, Databases, Computer Systems, Computer Networking, Artificial Intelligence, Machine Learning, Linear Algebra

Technical Skills

Languages: Java, Python, TypeScript, SQL, JavaScript, C, C++, HTML/CSS
Frameworks: Spring Boot, Flask, React.js, Node.js, Kafka
Libraries: TensorFlow, PyTorch, Pandas, NumPy, Scikit-learn
Developer Tools: Git, Docker, Jenkins, Gradle, AWS Lambda, AWS API Gateway, Google Cloud Run, Firebase, Postman, Jira

Experience

Incoming SDE Intern at Amazon — May 2026 – August 2026
Ads Response Prediction — Seattle, WA

Georgia Tech VIP - ConstructConnect LLM Research — August 2025 – December 2025
AI Research Intern — Atlanta, GA

- Built an **OCR, NER, LLM** pipeline in Python to process 80% of HVAC construction spec books, extracting structured data (manufacturer, context, warranty, compliance, components) with 91% accuracy across teams.
- Implemented a Retrieval-Augmented Generation (RAG) pipeline, turning chunking from NER terms into vectorized embeddings, enabling targeted context retrieval around manufacturer mentions; reduced boolean token usage by 60%.
- Created a provider-agnostic wrapper via Google Colab compatible with multiple LLM providers, exporting CSV, Excel, and JSON formats; **reduced manual review time by 85%** for downstream analysts at ConstructConnect.

Morgan Stanley — June 2025 – August 2025
Software Engineer Intern — Alpharetta, GA

- Built a reusable Java Spring testing framework that simulated 100+ of Reno's daily business events, enabling end-to-end validation of new features and reducing post-release defects.
- Automated end-to-end testing with a Jenkins CI/CD pipeline, cutting execution time from **4 hours to 45 minutes**.
- Developed parameterized shell & Java scripts for daily events (Kafka messages, cron jobs, file operations), increasing test coverage and catching integration defects early.
- Standardized framework use across business units, recording processes and training 40 engineers for faster adoption.

SimpliEarn — January 2025 – Present
Project Lead — Atlanta, GA

- Built a TypeScript front end and API layer that transforms earnings-calls into structured insights, reducing manual research time for investors.
- Implemented 3 Python microservices (stock-data generation, LLM chat, summarization) with FastAPI, containerized them with Docker, and deployed them as independent Google Cloud Run services with auto-scaling.
- Hosted Next.js frontend on Vercel with CI/CD-enabled API gateway calling serverless Cloud Run Python services, **reducing end-to-end response times by 40%**.
- Built a LangChain-based pipeline to generate GPT4 embeddings for 200 earnings-call transcripts and indexed them into Pinecone to enable low-latency semantic search across companies.

Projects

AWS GenAI Innovate Hackathon — October 2025

- Won **2nd place** in the AWS GenAI Innovate event by building Biblios, an AI-driven literature-analysis tool leveraging Python, AWS Bedrock, and NLP embeddings to cluster research papers and identify knowledge gaps.
- Developed Flask/FastAPI APIs delivering LLM-based summaries and embeddings for real-time research queries.
- Designed interactive visualizations using JavaScript and D3.js for clustering results and trend analysis dashboards.

Virtual Memory System Simulator — March 2025

- Implemented a virtual memory system simulator with page table management, context switching, and memory access translation, ensuring accurate replication of OS behavior.
- Designed and implemented page fault handler with FIFO and Approximate LRU page replacement, reducing simulated memory access latency and improving system throughput.
- Tracked memory system performance via AMAT, page faults, writebacks, and access statistics.

Morgan Stanley Code to Give Hackathon — September 2024

- Secured **2nd place** overall by developing a full-stack platform for personalized job recommendations, leveraging BERT transformers to generate contextual embeddings & Cosine Similarity for matching.
- Leveraged LLM-based TextKernel API for automated parsing of various file formats to store parsed resumes in AWS S3.

MLB Player Performance Machine Learning Model — April 2024 – June 2024

- Developed a machine learning pipeline in Python to predict next-season Wins Above Replacement for MLB players using pybaseball, scikit-learn, and Ridge Regression with sequential feature selection.
- Built data preprocessing workflows to clean and transform 20 years of batting stats (10K+ player-seasons), incorporating time-series backtesting and advanced features (player trends & correlation metrics), achieving RMSE $\approx$ .65.