

Sam Agarwal

📍 New York, USA

✉ samagarwal082@gmail.com

🌐 [-samarth-agarwal](#)

🔗 [samarth70](#)

📁 [Portfolio](#)

PROFESSIONAL SUMMARY

AI Software Engineer with expertise in building production-ready LLM systems, multi-agent architectures, and edge-native AI applications. Proven experience deploying scalable AI solutions using RAG, LoRA fine-tuning, and agentic workflows. Combines strong software engineering fundamentals (Java/Spring Boot, Python, React) with advanced ML deployment techniques to deliver cost-efficient, high-performance AI systems. Currently pursuing MS in Data Science at Stony Brook University.

EXPERIENCE

Generative AI Engineer Intern, ARMA AI Labs

June 2025 - Aug 2025

- Engineered a DeepEval-powered evaluation framework leveraging 7 metrics (BLEU, BERTScore, GEval, AnswerRelevancy, Faithfulness) with custom normalization, resulting in objective prompt comparison across LLMs.
- Revealed GPT-4's **96%** factual accuracy and Gemini's **40x** cost-efficiency by benchmarking models across 20+ articles, resulting in data-driven routing strategies for domain-specific generation.
- Optimized content delivery by delivering a production-ready dynamic routing pipeline, resulting in balanced performance across accuracy, latency, and cost requirements.

Senior Associate Software Developer (Salesforce), Genpact

July 2022 - July 2024

- Reduced manual data entry by **60%** for enterprise clients (GE, Huntsman) by engineering scalable Salesforce CRM architectures leveraging Apex triggers and Process Builder, resulting in enforced data integrity.
- Ensured low-latency data consistency globally by designing CORA workflow automation and integrating Salesforce with ERP systems via SFTP and Spring Boot microservices, resulting in synchronized pricing and orders.
- Enhanced component stability by progressing from Intern to Senior Associate and leveraging LWC, SOQL queries, and code reviews, resulting in resolved production defects and improved CRM visualization.

Machine Learning Engineer Intern, TannMann Foundation

June 2021 - Aug 2021

- Achieved **96%** validation accuracy in brand logo detection by optimizing a TensorFlow pipeline leveraging data augmentation and backbone freezing, resulting in mitigated overfitting.

TECHNICAL SKILLS

AI/ML & LLMs: PyTorch, TensorFlow, LangChain, RAG, LoRA Fine-tuning, MCP, DeepEval, Prompt Eng., Agentic Workflows, LLM-as-Judge

Languages & Backend: Python, Java, C++, SQL, Spring Boot, FastAPI, Flask, REST APIs, Hibernate, Apex

Data & Vector DBs: ChromaDB, Pinecone, Pandas, NumPy, Spark, ETL, AWS, Cloudflare Workers, Docker, Git

Frontend & Tools: React, Tailwind CSS, Vite, Pyodide/WASM, Hugging Face, Salesforce CRM, LWC, Visualforce

PROJECTS

ParamMem: Parametric Reflective Memory for LLM Agents [LoRA Fine-tuning, Llama-3, Groq API, ChromaDB, HumanEval/GSM8K]

- Mitigated reflection collapse and reduced inference latency by **94%** (45.2s to 2.8s) by offloading iterative reasoning to a parametric memory module powered by LoRA and Groq API.
- Architected ParamAgent-plus hybrid framework, leveraging ChromaDB retrieval to reduce prompt token footprint by **45%** while increasing task success rates (HumanEval **+11.4%**, GSM8K **+12.5%**).

PrepGenie: LLM Interview Coach [Pyodide/WASM, Recursive AI Failover, LLM-as-Judge, Resume-Aware RAG, Edge-Native]

- Reduced average TTFB from 1.2s to 120ms and cold starts from 5s to under 100ms by developing a Pyodide/WASM-optimized edge platform leveraging distributed SQL and async request handling.
- Ensured **99.9%** uptime while generating contextual questions with real-time scoring by engineering a recursive AI failover chain (Gemini → Groq → Mistral) leveraging pypdf and LLM-as-Judge.

Autonomous AI Penetration Testing Framework [React, Flask, Groq/Gemini, Multi-Agent, Nmap/Shodan, Docker, Cloudflare]

- Reduced security audit times from days to minutes by architecting a multi-agent AI framework leveraging 4 specialized agents with Groq/Gemini LLMs, resulting in rapid vulnerability detection with **99.9%** uptime.
- Achieved professional-grade accuracy by integrating Nmap/Shodan with automated CVE mapping while enforcing strict ethical guardrails, resulting in compliant, production-ready security assessments.

Manifesto Explainer: Political Party Analysis [NLP, Sentiment Analysis, Word Clouds, Heroku, Auto-Translation]

- Enabled unbiased political manifesto analysis by developing an NLP approach leveraging sentiment/subjectivity analysis and word cloud visualizations, resulting in comprehensive perspectives on party goals.
- Enhanced accessibility by implementing Party Symbol Detection and Auto-Translation features leveraging Heroku deployment, resulting in seamless multilingual political content interpretation.

EDUCATION

Stony Brook University SUNY, Master of Science in Data Science

Aug 2024 - May 2026 | New York, USA

Statistical Learning, Data Management, Big Data Algorithms & Networks, Statistical Computing, HCI, Frontiers in LLMs

Amity University, Bachelor of Technology in Computer Science & Engineering

June 2018 - June 2022

DBMS, Operating Systems, Data Structures, OOP, Java, Theory of Computation, Compiler Construction, Big Data, Cyber Security

ACHIEVEMENTS

[Salesforce AI Associate](#) | [SAP Generative AI Developer](#) | [Cisco CCNA Routing & Switching](#) | [ML in Production \(DeepLearning.ai\)](#) | [ML with Python \(IBM\)](#) | [AWS Solutions](#) | [BCG Data Science](#) | [KPMG Analytics](#) | [Goldman Sachs SWE](#) | [JP Morgan SWE](#) | [GSolve.ai ML Hackathon 2023 Winner](#) | [Geoffrey Hinton Fellowship Hackathon](#) | [Pneumonia Detection \(Top 50/2000+\)](#) | [Univ.ai \(Rank 123\)](#) | [Dare in Reality \(Rank 159\)](#) | [UHack \(Rank 160\)](#) | [Wipro \(Rank 284\)](#)