Samarth Agarwal

(332)250-9030 | samarthagarwal26@yahoo.com

American/Singapore PR | https://github.com/Samarth26

Academic Qualifications

New York University, New York Aug 2024 - May 2026

Master of Science, Data Science

- Optimization and Linear Algebra, Machine Learning, Big Data, Deep Learning, Natural Language Processing

Nanyang Technological University, Singapore Aug 2020 - May 2024

Bachelor of Science (Honors), Data Science and Artificial Intelligence

- Information Retrieval, Intelligent Agents, Computer Vision, Developing Data Products, Simulation Techniques in Finance

Projects

Agentic AI Pipeline for Freight Document Parsing & Email Automation {OCR, LangChain} Sept 2025 – Dec 2025

- Built an agentic OCR–LLM pipeline for freight documents using schema-constrained GPT parsing, deterministic OCR grounding, and validation agents, reducing mislabel rate from 14.2% to 1.8% and CER from 9.5% to 2.5%.
- Developed a safety-aware email automation system with hybrid intent classification and abstention logic, achieving 70% ready-to-send replies, 8% hallucination rate, and explicit human review for 19% uncertain cases.

Self-Supervised Representation Learning {pytorch} Oct 2025 – Dec 2025

- Trained a ResNet-101 encoder from scratch using self-supervised learning (SimCLR, BYOL, MoCo), systematically analyzing convergence behavior and identifying MoCo as the most stable and scalable objective under limited compute.
- Learned representations from 600k+ unlabeled images and achieved 59.9% top-1 accuracy on mini-imagenet and 34.5% on CUB-200 using a frozen encoder with a 30-epoch linear probe and tuned hyperparameters.

Prompt-Side Activation Suppression in LLMs {dreamy} Oct 2025 – Dec 2025

- Built an inverted Evolutionary Prompt Optimization (EPO) pipeline to minimize targeted logits, neuron activations, and residual-stream directions while preserving prompt fluency via Pareto (target vs. cross-entropy) optimization.
- Conducted layerwise sensitivity sweeps on microsoft/phi-2 to localize features and showed prompt-only search beats random prompting and dataset-scan baselines for driving targets toward near-zero activation.

Movie Recommender {pyspark, scipy} Mar 2025 – May 2025

- Developed and evaluated recommendation engines using Apache Spark's ALS matrix factorization and multiple temporal data-splitting strategies (leave-one-out, rolling window), achieving significant RMSE reductions compared to baseline models.
- Implemented scalable similarity analysis with MinHashLSH to efficiently compute user–user similarities across millions of interactions, validating results through Pearson correlation and identifying high-quality "movie twin" pairs.

ML Approach to Crowd Trajectory Prediction, {PyTorch, Java, C++} Jun 2023 - May 2024

- Designed and deployed a multilayer perceptron model in Java and C++ for modeling crowd behavior using video data which performed better than the shortest path algorithm and the Social Force Model by achieved a density error of 699.54
- Customized MASON Simulation package to fit the simulation parameters and testing environment to conduct simulation studies.

Professional Experience

Federal Reserve Board, Graduate Data Intern Jan 2025 – Sep 2025

- Architected and implemented a Python-based entity matching pipeline, leveraging feature embeddings, string similarity metrics, and custom feature engineering to enable fuzzy name matching across a 30M+ record corpus.
- Conducted exploratory data analysis on a historical record corpus to segment it into three distinct entity clusters, sharpening our pattern-recognition focus and improving downstream matching efficiency
- Applied MinHash & LSH to approximate Jaccard similarity, slashing per-record candidate sets from 3 million to ~10 (over 99.999% reduction) and enabling end-to-end matching at scale

Tata Consultancy Services, Machine Learning Intern Jun 2023 - Aug 2023

- Developed a marketing chatbot that streamlines content & creatives creation for Banks and Financial services using Langchain, vector stores, and stable diffusion.
- Through data collection and prompt engineering, tuned the LLM to be geared towards producing creative captions and images for social media.

Haleon, Business Analyst Intern Jun 2022 - Jan 2023

- Led the Innovation Dashboard project, overseeing the development of a PowerBi dashbord that projected a reduction in labor costs by over 80%, and delivered the project two months ahead of schedule (6 months to 4 months).
- Performed data wrangling, validation and analyzed structured data using Excel, PowerQuery, DAX and Python to uncover additional metrics which allowed us to exceed stakeholder requirements.

Skills

Software Skills: Python, R, SQL, C++, Java, JavaScript, HTML/CSS, React, Angular, GIT, DAX, LaTeX

Software Applications and Frameworks: PyTorch, scikit-learn, XGBoost, LGBM, Optuna, pandas, NumPy, Langchain, Firebase, Flutter, MongoDB, MASON, PowerBi, Microsoft Office, Bitbucket, GitHub, MATLAB