SAMARTH AGARWAL

+1 (332) 250-9030 | samarthagarwal262000@gmail.com | New York, NY, USA | Citizenship: American / Singapore PR |
linkedin.com/in/samarth-agarwal-0499b1181/ | github.com/Samarth26 | samarth26.github.io/

PROFESSIONAL SUMMARY

Applied AI Scientist (MS Data Science) specializing in **representation learning, embedding models, and personalization systems**. Experienced in self-supervised learning, large-scale entity matching, recommender systems, and agentic LLM pipelines. Strong focus on translating research ideas into scalable production ML systems.

EDUCATION

New York University — **August 2024 - May 2026**
Master's, Data Science — *GPA: 3.89*
- Optimization and Linear Algebra, Machine Learning, Big Data, Deep Learning, Natural Language Processing

Nanyang Technological University (NTU) — **August 2020 - May 2024**
Bachelor's, Data Science
- Information Retrieval, Intelligent Agents, Computer Vision, Developing Data Products, Simulation Techniques in Finance

PROFESSIONAL EXPERIENCE

The Federal Reserve System — **Washington, DC, USA**
Graduate Data Intern — *January 2025 - September 2025*
- Addressed an unsupervised entity resolution problem involving inconsistent and non-canonical names across three datasets, where 30M+ records made exhaustive pairwise string comparisons computationally infeasible and ruled out supervised classification due to the absence of ground truth.
- Formulated entity matching as a pattern-recognition and similarity-scoring task, relying primarily on string-based signals to extract discriminative features and assign match confidence under noisy and imperfect auxiliary attributes.
- Implemented a two-stage matching architecture: first applying MinHash with Locality-Sensitive Hashing to approximate Jaccard similarity and aggressively prune the search space (>99.999% reduction), followed by precise string similarity scoring (e.g., Jaro–Winkler) to produce high-confidence entity alignments at scale.

Haleon — **Singapore**
Business Analyst Intern — *June 2022 - January 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.

PROJECTS & OUTSIDE EXPERIENCE

Agentic AI Pipeline for Freight Document Parsing & Email Automation — **New York, NY, USA**
September 2025 - December 2025
- Built an agentic OCR–LLM pipeline for freight documents with langgraph 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

Prompt-Side Activation Suppression in LLMs — **New York, NY, USA**
October 2025 - December 2025
- Investigated the problem of controlling and suppressing specific internal model activations (target logits, neurons, and residual-stream directions) using only prompt-level interventions, motivated by interpretability and safety settings where model weights and training data are inaccessible
- Developed an inverted Evolutionary Prompt Optimization (EPO) pipeline, applying Pareto optimization between target-activation objectives and cross-entropy constraints, and demonstrated via layerwise sensitivity sweeps on Phi-2 that prompt-only search consistently drives target activations toward near-zero, outperforming random prompting and dataset-scan baselines

Self-Supervised Representation Learning — **New York, NY, USA**
October 2025 - December 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 on the downstream task of image classifcation

Movie Recommender — **New York, NY, USA**
March 2025 - May 2025
- Developed and evaluated recommendation engines on the imdb movies dataset 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" pair

ML Approach to Crowd Trajectory Prediction — **Singapore**
June 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

SKILLS

Skills: AngularJS, React.js, HTML/CSS, JavaScript, Java, MySQL, LangChain, Flutter, Firebase, MATLAB, Python, Pytorch, Git, Excel/Numbers/Sheets, Business Analytics, MongoDB, C/C++, Data Science, Data Structures & Algorithms, Google Cloud Platform, NoSQL, Node.js, Pandas, Power BI, R, SQL, NumPy, Optimization, Deep Learning, Fuzzy Matching, Natural Language Processing (NLP), Computer Vision, DAX, LaTeX, XGBoost, LGBM, Optuna, MASON, Microsoft Office, Bitbucket, Data Wrangling, Feature Engineering, Exploratory Data Analysis, Hypothesis Testing, Simulation Techniques, Collaborative Filtering, Recommendation Systems, Image Classification, Transfer Learning, Hyperparameter Tuning, Prompt Engineering, Pattern Recognition, Scikit-learn, Machine Learning, Data Analysis, Entity Matching, Project Management, Creative Problem Solving, Team Collaboration, Communication, Tensorflow