# Samir Char

sc.samirchar@gmail.com | www.samirchar.com | www.linkedin.com/in/samir-char | www.github.com/samirchar
Boston, Massachusetts | 646-262-8466

## Research Interests

Advance self-supervised multimodal learning to build foundation models that generalize across domains. I am also interested in artificial intelligence applications in biomedicine and scientific discovery (AI4Science), where unlabeled, diverse data abounds.

## Education

**Columbia University** New York, NY, US
**Master of Science in Data Science** (GPA: 4.08/4.0) Jan. 2021 – May 2022

- Relevant Courses: Machine Learning, Deep Learning, Advanced Deep Learning, Natural Language

**Universidad de Los Andes** Bogota, Bogota D.C., CO
**Bachelor of Science in Electrical Engineering** (GPA: 4.1/5.0) Aug. 2013 – Dec. 2018

- Relevant Courses: Intelligent Analysis of Signals and Systems, Signals, Optimization

## Publications & Preprints

Equal contribution indicated with *

3. Char, S., Corley, N., Alamdari, S., Yang, K.K., Amini, A.P. (2025). ProtNote: a multimodal method for protein–function annotation. *Bioinformatics*. DOI: doi.org/10.1093/bioinformatics/btaf170
2. Yang, K.K.*, Alamdari, S.*, Lee, A.J.*, Kaymak-Loveless, K., Char, S., Brixi, G., Domingo-Enrich, C., Wang, C., Lyu, S., Fusi, N., Tenenholtz, N., Amini, A.P. (2025). The Dayhoff Atlas: scaling sequence diversity Improves Protein Design. *bioRxiv*. DOI: doi.org/10.1101/2025.07.21.665991
1. Upadhyay, V. P., Modi, S.*, Gupta, S. A.*, and Char, S*. (2024). Exploring Few-Shot Performance of Self-Supervised Visual Representations. *International Conference on Emerging Trends in Networks and Computer Communications (ETNCC).* DOI: 10.1109/ETNCC63262.2024.10767558

## Research Experience

**Microsoft, BiomedicalML Lab (Nicolo Fusi's Group)** Boston, MA, US
**Independent Researcher** Jul. 2023 – Present (~15 hrs/week)

- Collaborating with Microsoft Research on AI methods for biological discovery and engineering. Involves distributed (e.g., FSDP, DDP), multimodal training with Transformers, CNNs, and LLMs. Conducted research outside main role. Advisors: Dr. Kevin K Yang and Dr. Ava Amini. Outcomes: 1 publication, 1 preprint.

**Columbia University, Data Science Institute** New York, NY, US
**Data Science Institute Scholar** Jan. 2022 – Jul. 2022

- Selected for the competitive DSI Scholars program to advance research at the intersection of machine learning and clinical neurology. Advisors: Dr. James Noble and Dr. Olajide Williams.
- Spearheaded a smartphone-based multimodal (video, speech, tabular) stroke prediction model using computer vision (CNN) and machine learning. Achieved 89% sensitivity and 58% specificity, making us finalists in Columbia's BiomedX competition.

**Universidad de Los Andes, Department of Electrical Engineering** Bogota, Bogota D.C., CO
**Undergraduate Researcher** Jan. 2017 – Dec. 2017

- Thesis: Pioneered a machine learning ensemble (XGBoost, Random Forest) to predict patients' hospital length of stay at admission to improve how low-income hospitals manage workforce, facilities, and resources. Algorithm surpassed doctors' accuracy by 15% (absolute). Advisor: Dr. Luis Felipe Giraldo. Link

## Industry Experience

**Microsoft, Azure Core** Remote – Boston, MA, US
**Applied Scientist 2** Jun. 2022 – Present

- Created a hierarchical forecast and Monte Carlo simulation to detect customer capacity risks over a six-month horizon. Tool deployed across 65 regions, identifying 400K+ vCore gaps that drove procurement of new servers to mitigate global risk.
- Utilized causal inference (Double ML) to estimate the influence of customer experience managers on customers' capacity experience. Found that customer management reduces backlogged quota requests by 20% and capacity failures by 0.3%
- Developed a recommendation engine to prioritize global quota requests, enabling triage of over 2M vCore requests to date.
- Improved XGBoost model to forecast Azure global quota requests, helping allocate capacity across Data Centers globally.

- Developed a forecasting ensemble (ARIMA + ETS + regression) for Azure support tickets, resulting in 20% improvement from baseline (absolute; MAPE). Forecast used to estimate annual staff and budget needs is approximately $40M/year.
- Led product analytics initiative, defined metrics, drove cross-team alignment, and launched a Power BI that cut manual queries by 92% and time-to-insight by 40%

**Microsoft, Azure Core** Remote – New York, NY, US
**Applied Scientist Intern** Jul. 2021 – Sep. 2021

- Found ways to improve an Azure quota ticket forecasting model by using new data, machine learning, and causal inference.
- Analyzed Azure Data Explorer usage data to identify top customers, usage patterns, projected growth, and churn behavior.

**Mercado Libre** Bogota, Bogota D.C., CO
**Applied Scientist** Jun. 2020 – Dec. 2020

- Enhanced a boosting model multi-warehouse demand forecasting to ensure accurate inventory plans by stocking items customers want in various regions; collaborated with 10 Applied Scientists to meet commercial needs.

**Grupodot** Bogota, Bogota D.C., CO
**Applied Scientist** Oct. 2018 – Jun. 2020

- Crafted a deep learning model (RNN) to forecast fuel sales for 200+ gas stations of a major oil & gas company, boosting the client's model MAPE by 11% (absolute).
- Built a gradient boosting churn model for Latin America's largest telecom; attained a 5.2 uplift on 200K monthly customers.
- Helped develop a credit default prediction model for a major Colombian bank to secure resources in case of client default.
- Built Python web scrapers and social media analytics tool using Google APIs for comment clustering and influencer scoring.

**Young and Rubicam** Bogota, Bogota D.C., CO
**Junior Applied Scientist** Jun. 2018 – Oct. 2018

- Automated a Machine Learning tool to measure social media success using Python, reducing manual tasks by 85%.

**IBM** Bogota, Bogota D.C., CO
**Intern** Jan. 2018 – Jun. 2018

- Designed a machine learning workshop given to more than 150 entrepreneurs, Master students and undergraduates.
- Recruited disruptive technology-based startups; identified IBM Cloud technologies to solve business needs.

## Teaching Experience

**Microsoft, Azure Core** Remote – Boston, MA, US
**Lecturer** Jun. 2024

- Designed and presented lectures on statistics for Microsoft's Azure Learning Academy.

**Columbia University, Department of Computer Science** New York, NY, US
**Lecturer, Advanced Topics in Deep Learning Seminar** Mar. 2022

- Designed and taught lectures on self-supervised learning models and object detection architectures to ~25 students.

**Grupodot** Santiago, Santiago Metropolitan Region, CL
**Instructor** Aug. 2019

- Developed and delivered a week-long machine learning & Google Cloud workshop for 50 engineers.

**Universidad de Los Andes, Department of Electrical Engineering** Bogota, Bogota D.C., CO
**Teaching Assistant, Control Systems** Jan. 2017 – Jun. 2017

- Lectured on course material and led review sessions. Wrote, facilitated, and graded exams.

## Talks & Presentations

**Remando en Arequipe Podcast** | Virtual Aug. 2025

- Invited guest – Discussed career path, AI and biomedical applications, reaching Latin American audience.

**Universidad de Los Andes, M.S. in Artificial Intelligence** | Virtual Jun. 2025

- Invited talk – ProtNote: a multimodal method for protein-function annotation.

**Universidad de Los Andes, M.S. in Artificial Intelligence** | Virtual Dec. 2024

- Invited talk – ProtNote: a multimodal method for protein-function annotation.

## Leadership & Outreach

**Microsoft HOLA (Hispanic and Latinx Organization of Leaders in Action)** | Boston, MA, US Jan. 2024 – Present

- Speaking & Outreach – support Hispanic/Latinx community growth; promote diversity; review scholarship applications.

**Pro-bono Mentor** | Virtual Jun. 2021 – Present

- Mentorship – mentor students and early-career professionals pursuing AI careers through one-on-one conversations.

**Microsoft** | Remote – Boston, MA, US Jun. 2024 – Present

- Mentorship – mentor early-career employees on technical and career development.

**Columbia University** | Virtual May 2025

- Volunteer – conducted mock technical and behavioral interviews for M.S. in Data Science students.

## Skills

- **ML & AI**: Deep Learning, Machine Learning, Time Series Forecasting
- **Frameworks & Libraries**: PyTorch, TensorFlow, XGBoost, Hugging Face, PyTorch Lightning, scikit-learn, NumPy, Pandas
- **Distributed Training**: Fully Sharded Data Parallel (FSDP), Distributed Data Parallel (DDP)
- **Programming & Systems**: Python, SQL, R, UNIX, Linux, JavaScript
- **ML Ops & Data Engineering**: Weights & Biases, Apache Spark (PySpark)
- **Cloud**: Microsoft Azure, Google Cloud Platform, Amazon Web Services
- **Databases**: Azure Data Explorer (Kusto), BigQuery, MySQL
- **Data Visualization**: Matplotlib, Seaborn, Power BI
- **Languages**: English (fluent), Spanish (native)

## Personal Interests

- Running (completed four half marathons), weightlifting, martial arts (karate black belt), books, fine dining