Gaurav Sett

Contact | gauravsett@icloud.com | linkedin.com/in/gauravsett | gauravsett.com

Education

Georgia Institute of Technology **Atlanta, GA**

Master of Science in Computer Science (4.0 GPA) *2023-01 to 2023-08*

- Machine Learning Specialization — Courses in Deep Learning, Natural Language Processing, Advanced Algorithms

Georgia Institute of Technology **Atlanta, GA**

Bachelor of Science in Computer Science (3.9 GPA) *2019-08 to 2022-12*

- Artificial Intelligence Thread — Took courses in Machine Learning, Artificial Intelligence, Automata and Complexity
- Human-Computer Interaction Thread — Took courses in Cognitive Science, Social Psychology, Educational Technology
- Honors Program Student — Awarded distinction in research by completion of undergraduate thesis

Work

Center for AI Safety **San Francisco, CA**

Fundraising Contractor *2023-09 to Present*

- Conducting donor research and analysis to identify potential supporters for leading AI safety non-profit

Washington Post **Washington, DC**

Data Science Intern *2022-05 to 2022-08*

- Categorized article topics in breaking news stream with LSTM and BERT models using PyTorch
- Created API to handle continuous collection, processing, and storage of articles using AWS ECS, Lambda, and S3
- Visualized breaking news topics and API performance on dashboard for journalists using Datadog

Federal Reserve Board **Washington, DC**

Economics Research Intern *2021-05 to 2021-08*

- Developed text analysis tool regularly used in policy meetings to analyze corporate earnings calls using SpaCy
- Automated collection, processing, and storage of 200K documents; used parallelization to improve speed 8x with Dask
- Presented to economists on NLP techniques such as word vectors and LDA topic modeling with interactive Django app

Georgia Tech Research Institute **Atlanta, GA**

Undergraduate Research Intern *2020-05 to 2020-12*

- Created classification models identifying conspiratorial anti-vax Reddit comments with 80% accuracy using SciKit-Learn
- Analyzed the partisan differences in tweets about COVID-19 topics from members of Congress using SpaCy
- Built GUI application enabling social scientists to collect and analyze Twitter API data using PyQT5

Leadership

Supervised Program for Alignment Research **Remote**

Founder & Program Manager *2023-07 to Present*

- Launched program facilitating AI safety research projects for 15 advisors and 80 students at 12 universities
- Recruiting participants, managing applications, and organizing 18 projects in alignment, engineering, and policy

AI Safety Initiative at Georgia Tech **Atlanta, GA**

Founder & Managing Director *2022-08 to Present*

- Founded group hosting events, seminars, bootcamps, and research projects in AI alignment and governance
- Managing programs for dozens of students, leading team operations, and supporting career transitions
- Open Philanthropy University Organizer Fellowship — Awarded $7,100 grant to support work as group organizer
- Open Philanthropy Group Support Grant — Secured $5,000 grant to support research and education programs

Research

EleutherAI **Remote**

AI Alignment Projects *2023-05 to Present*

- Adversarial Robustness — Trained language model on reverse order data to identify potential jailbreak inputs
- Unpaired Image Generation — Produced images from text with minimal paired training data using shared latent space

Georgia Tech College of Computing **Atlanta, GA**

Graduate Projects *2023-01 to 2023-08*

- PhilBench — Developed value alignment benchmark for reward modeling methods using text corpora
- Conference Presentation — Sherri L Conklin, Gaurav Sett (2023). AI Safety, Governance, and Alignment Tutorial. *2023 IEEE International Symposium on Ethics in Engineering, Science, and Technology (ETHICS)*
- Language Model Analysis — Measured the ability of GPT-3 to rationalize moral judgements in a chatbot setting
- Co-Citation Prediction — Measured likelihood of paper co-citation using graph neural networks and BERT

Undergraduate Projects *2019-08 to 2022-12*

- Gentrification Forecasting — Built transformer model forecasting of Atlanta gentrification with tax data using PyTorch
- Expert Models Review — Wrote literature review on potential development of chatbots with subject matter expertise
- Portfolio Clustering — Analyzed investment portfolios to identify clusters of asset price time-series using K-means
- Employment Review — Wrote literature review on the likely effects of AI on wages and employment over this century
- Social Media Review — Conducted literature review on how social media amplifies misinformation in public discourse

Teaching

Introduction to Cognitive Science **Atlanta, GA**

Graduate Teaching Assistant *Spring 2023*

Graduate Teaching Assistant *Summer 2023*

Skills

Machine Learning | Python | R | SQL | PyTorch | TensorFlow | HuggingFace | SciKit-Learn | Pandas | NumPy | Matplotlib

Software Engineering | Java | C | JavaScript | HTML/CSS | Git | APIs | AWS | React | Django | Agile