Seraphina Goldfarb-Tarrant

seraphinatarrant.github.io, seraphina@cohere.com, +1 415.683.7627

SUMMARY

9 years of experience working in tech, plus a PhD in NLP. This includes:
- NLP (Natural Language Processing) research in: Fairness, Causality, Unsupervised Learning, Natural Language Generation, Information Retrieval, Multilinguality.
- 5 first author NLP publications at top tier conferences, 1 second author
- 4.5 years at Google as a global Product Manager & programmer in adTech. Joined a recently acquired team of five, helped grow the product to $1 billion annual revenue and 1 million queries-per-second .
- 1.5 years at sea teaching disadvantaged teens teamwork via sailing.

EDUCATION

U. of Edinburgh, Edinburgh — *PhD Candidate (awaiting viva) in Informatics: Institute for Language, Cognition, and Computation*

September 2019 - Feb 2024 (expected), advised by Adam Lopez

Thesis: Fairness in Transfer Learning.

U. of Washington, Seattle — *MSc in NLP/Computational Linguistics*

September 2017 - June 2019 , GPA: 3.97

Projects: classification, sequence-tagging, language models, linguistics, parsing, topic models, automatic summarization, information retrieval

UCLA, Los Angeles — *BA Ancient Greek, Minor in Film*
Grad date: September 2010, GPA 3.98

UCLA Honors Program, summa cum laude, Phi Beta Kappa, 5 scholarships

PROFESSIONAL EXPERIENCE

Cohere, London – *Head of Safety*

May 2023 - Present

Head of all Safety efforts for our LLM, and Modelling Tech Lead

Meta (FAIR), London — *Research Scientist Intern*

Sept 2022 - April 2023

Fairness and Interpretability for Retrieval Augmented QA systems.

Amazon, Barcelona — *Research Scientist Intern*

August 2021 - Jan 2022

Multilingual transfer learning for sentiment analysis.

Bayes Centre for DataScience, UK — *Lead NLP Engineer*

November 2019 - Dec 2022

In collaboration with the Gates Foundation, created a multilingual event and entity extraction ML system to map the spread of livestock diseases across 12 African countries. Presented at the UN for LD4D in Feb 2020.

Information Sciences Institute, LA — *Research Engineer*

July 2018 - June 2020

Research in Narrative Generation, Human-AI interaction & DARPA NLP projects. Build novel state-of-the-art systems in PyTorch. Responsible

Programming Languages

Fluent: Python & Django

Conversational: C++, SQL

Natural Languages

Conversational: Japanese, Spanish, Ancient Greek

Skills

Machine Learning

Neural Networks

Jira, Git, Latex, Balsamiq, Sketch, Alexa Platform

PROFESSIONAL SERVICE

Reviewer (2020 - present):
ACL, NAACL, EMNLP, ARR

Ethics Committee (2022 - present):
NEURIPS, EMNLP, EACL

Workshop Organiser:
Gender Bias in NLP (ACL 2023), RepL4NLP (ACL 2023)

AWARDS

- Apple AI Fellow Nomination (5 students/year)
- First-place Alexa social bot (restaurant-recommender) in Amazon competition

PROJECTS

Interactive NLP

Built an Interactive Neural Story Writing system: video

Human-AI Interfaces

Collaboration with performance artists to investigate how the public views and interacts with AI systems. First shows premiered in Amsterdam and NYC (2019), in collaboration with the VALES project: link

Excelsior Trust (non-profit)

Work on *Excelsior*, a historic wooden 23m ship in the North Sea & Baltic, which offers team building for teens & recovering addicts.

Began as a volunteer, subsequently hired as relief Boatswain & First Mate, then elected to the board. 15,000 nautical miles.

Digital Mapping of the

for research experiments & authoring works for publication.

Google, Tokyo, NYC, Shanghai — *Product Manager, AdTech*

July 2012 - Oct 2015

- Launched main product in AU/NZ, Japan, China by increasing local partner integrations by 300%, and driving localization and language detection/categorization. Developed the Asia-Pacific region to a revenue growth of 212% YoY ($11M to $34M).

- PM for global relationship with Yahoo!; eliminated primary source of on-call incidents.

- PM for global Machine Learning bidding algorithm improvements, influencing 30% of revenue, and for anti-Malware features.

Google, NYC— *Technical Account Manager, AdTech*

January 2011 - June 2012

- Designed custom solutions in python/django for clients responsible for 40% of product revenue.

- Organised QA (technical and user) for all releases and for migration from Amazon ec2 to Google infrastructure.

Ancient World

Developed 3D reconstructions of Ancient Rome throughout time, to integrate a *time dimension* into Google Earth. As part of the UCLA Experiential Technologies Center, in 2010. (The Google Earth API has now been turned down, but more work done by the same lab can be found here)

PUBLICATIONS

MultiContrievers: Analysis of Dense Retrieval Representations — *(Under Submission)*
Seraphina Goldfarb-Tarrant, Pedro Rodriguez, Jane Dwivedi-Yu, Patrick Lewis

Cross-lingual Transfer Can Worsen Bias in Low-Resource Sentiment Analysis — EMNLP 2023
Seraphina Goldfarb-Tarrant, Björn Ross, Adam Lopez
arXiv link

This Prompt is Measuring <MASK>: Evaluating Bias Evaluation in Language Models — ACL 2023
Seraphina Goldfarb-Tarrant, Eddie Ungless, Esma Balkir, Su Lin Blodgett
arXiv link

Bias Beyond English: Counterfactual Tests for Bias in Sentiment Analysis in Four Languages — ACL 2023
Seraphina Goldfarb-Tarrant, Adam Lopez, Roi Blanco and Diego Marcheggiani
arXiv link

How Gender Debiasing Affects Internal Model Representations, and Why It Matters — NAACL 2022
Hadas Orgad, **Seraphina Goldfarb-Tarrant**, Yonatan Belinkov
arXiv link

Intrinsic Bias Metrics Do Not Correlate with Application Bias — ACL 2021
Seraphina Goldfarb-Tarrant, R. Marchant, R. Muñoz Sanchez, Mugdha Pandya, Adam Lopez
arXiv link, video link

Content Planning for Neural Story Generation with Aristotelian Rescoring — EMNLP 2020
Seraphina Goldfarb-Tarrant, Tuhin Chakrabarty, Ralph Weischedel, Nanyun Peng
arXiv link, video link

Scaling Systematic Literature Reviews with Machine Learning Pipelines — SDP@EMNLP 2020
Seraphina Goldfarb-Tarrant, A. Robertson, J. Lazic, T. Tsouloufi, L. Donnison, K. Smyth
arXiv link

Plan, Write, and Revise: an Interactive System for Open-Domain Story Generation — NAACL 2019
Seraphina Goldfarb-Tarrant, Haining Feng, Nanyun Peng
arXiv link, demo video

PRESS & MEDIA

LLM Generative Red Teaming at DEFCON (BBC, 2023)

Gender Bias in Machine Translation for Danish (Tjekdet, 2022) (in Danish)

INVITED TALKS

Invited Talks:

- Speaker at WOAH (Workshop on Online Abuse and Harms) NAACL 2024
- 3 upcoming talks at Edinburgh, Aberdeen and Cambridge Universities, November and December 2023
- Speaker at **ARIAS Amsterdam** on AI and the Arts, September 2023
- Panel talk on Responsible Generative AI with **Salesforce Ventures x Dawn Capital,** September 2023
- Webinar on Responsible AI with **Five9** (here), July 2023
- Panel talk on limitations of LLMs at RepL4NLP, **ACL 2023,** with Yejin Choi, Swabha Swayamdipta, Samira Abnar
- *Interpretability for Retrieval Augmented Generation*, **Technion Israel Institute of Technology**, Jan 2023
- *Bias in Language Model Representations*, **NYU,** September 2022
- Panel talk at Gender Bias in NLP workshop, **ACL 2022,** with Kai-Wei Chang, Kellie Webster, and Mark Yatskar
- *Understanding and Applying Bias Metrics for NLP Systems*, **National Research Council Canada,** March 2022
- Panel talk at Generation, Evaluation, Metrics (GEM), **ACL 2021**, with Ehud Reiter, He He, and Hady Elsahar
- *Interpretability and Reproducibility Workshop*, **Information Sciences Institute**, Sept 2019

TEACHING

University of Edinburgh, Informatics:

Tutor:

- NLU+ (Natural Language Understanding)
- MLPR (Machine Learning Practical)

Develop Coursework (assignments, labs) for ANLP (Accelerated NLP)

Primary supervisor for Informatics MSc Student dissertations

UCLA:
Tutor (2008-2010):

- C++, Astronomy, Earth & Space Sciences, ESL Composition, Ancient Greek

PUBLIC PROFILES

Github: https://github.com/seraphinatarrant (active since 2017)
LinkedIn: https://www.linkedin.com/in/seraphinatarrant/
Personal Website: seraphinatarrant.github.io