Seraphina Goldfarb-Tarrant

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

SUMMARY

11 years of experience working in tech globally, and a PhD in NLP.
This includes:
- Leading the safety team for one of the leading foundation model developers
- NLP (Natural Language Processing) research in: Fairness, Causality, Unsupervised Learning, Natural Language Generation, Information Retrieval, Multilinguality.
- Many first author and last author NLP publications at top tier conferences,, with over a dozen speaking engagements including keynotes at *CL workshops, as well as at WEF Davos and on BBC4
- 4.5 years at Google as a global Product Manager & engineer in adTech. Joined a recently acquired team of five, helped grow the product to $1 billion annual revenue and 1 million queries-per-second. Launched the product in Asia-Pacific region.
- 1.5 years at sea teaching disadvantaged teens teamwork via sailing.

EDUCATION

U. of Edinburgh, Edinburgh — *PhD in Informatics: Institute for Language, Cognition, and Computation*

September 2019 - May 2024 , 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 AI Safety*

May 2023 - Present

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

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.

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

Board of Advisors:
- *Humane Intelligence Algorithmic Bias Bounty*
- *University of Edinburgh CDT in Responsible NLP*

Area Chair, Safety (2024 - present):
ICML, COLM, ARR

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

Workshop Organiser:
Gender Bias in NLP (ACL 2024), RepL4NLP (ACL 2024)

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

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

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 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— *Customer Solutions Engineer, 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.

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 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

MAPS: A Multilingual Benchmark for Global Agent Performance and Security
Omer Hofman, Jonathan Brokman, Oren Rachmil, Shamik Bose, Vikas Pahuja, Toshiya Shimizu, Trisha Starostina, Kelly Marchisio, **Seraphina Goldfarb-Tarrant**, Roman Vainshtein

Small Changes, Large Consequences: Analyzing the Allocational Fairness of LLMs in Hiring Contexts AACL 2025
Preethi Seshadri, Hongyu Chen, Sameer Singh, **Seraphina Goldfarb-Tarrant**

A Good Plan is Hard to Find: Aligning Models with Preferences is Misaligned with What Helps Users EMNLP 2025
Nishant Balepur, M Shu, Y Y Sung, **Seraphina Goldfarb-Tarrant**, S Feng, F Yang, R Rudinger, Jordan Lee Boyd-Graber

Safer or Luckier? LLMs as Safety Evaluators Are Not Robust to Artifacts ACL 2025
Hongyu Chen, **Seraphina Goldfarb-Tarrant**

The Multilingual Divide and Its Impact on Global AI Safety
Aidan Peppin, Julia Kreutzer, Alice Schoenauer Sebag, Kelly Marchisio, Beyza Ermis, John Dang, Samuel Cahyawijaya, Shivalika Singh, **Seraphina Goldfarb-Tarrant,** Viraat Aryabumi, Wei-Yin Ko, Ahmet Üstün, Matthias Gallé, Marzieh Fadaee, Sara Hooker

Scalpel vs. Hammer: GRPO Amplifies Existing Capabilities, SFT Replaces Them ICML Interpretability Workshop 2025
Neel Rajani, Aryo Pradipta Gema, **Seraphina Goldfarb-Tarrant**, Ivan Titov

Command A: An Enterprise-Ready Large Language Model
Team Cohere: (I was one of two primary authors collecting and editing contributions of 100+ person team)

Mix Data or Merge Models? Optimizing for Diverse Multi-Task Learning NeurIPS 2024 Workshop SafeGenAi
Aakanksha, Arash Ahmadian, **Seraphina Goldfarb-Tarrant**, Beyza Ermis, Marzieh Fadaee, Sara Hooker

The Multilingual Alignment Prism: Aligning Global and Local Preferences to Reduce Harm EMNLP 2024
Aakanksha, Arash Ahmadian, Beyza Ermis, **Seraphina Goldfarb-Tarrant**, Julia Kreutzer, Marzieh Fadaee, Sara Hooker

A SMART Mnemonic Sounds like "Glue Tonic": Mixing LLMs with Student Feedback EMNLP 2024
Nishant Balepur, M Shu, A Hoyle, Alison Robey, Shi Feng, **Seraphina Goldfarb-Tarrant,** Jordan Boyd-Graber

MultiContrievers: Analysis of Dense Retrieval Representations Blackbox NLP @EMNLP 2024
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

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

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

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

Intrinsic Bias Metrics Do Not Correlate with Application Bias (**recorded presentation**) ACL 2021
Seraphina Goldfarb-Tarrant, R. Marchant, R. Muñoz Sanchez, Mugdha Pandya, Adam Lopez

Content Planning for Neural Story Generation with Aristotelian Rescoring (**recorded presentation**) EMNLP 2020
Seraphina Goldfarb-Tarrant, Tuhin Chakrabarty, Ralph Weischedel, Nanyun Peng

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

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

PRESS & MEDIA

Royal Institution Christmas Lectures 2023, *The Truth About AI*, BBC4

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 **Chatham House** *Beyond the hype: The realities and risks of artificial intelligence today*, October 2025
- Invited talk on *Emerging Challenges in NLP Safety* at **Heriot Watt**, October 2025
- Safety Lecturer at **M2L (Mediterranean Machine Learning)** Summer School, September 2025
- Keynote Speaker at University of Edinburgh **UKRI CDT in NLP/SLT Joint Conference**, June 2024
- Keynote Speaker at WOAH (Workshop on Online Abuse and Harms), **NAACL 2024**
- AI Fringe Panel on the **AI Safety Summit**, at the British Library, June 2024
- Panel on *Strategies for Safer AI*, **NVIDIA GTC,** March 2024
- Balancing AI Bias panel at **Bloomberg London**, talk for International Women's Day, March 2024
- Steering LLM Safety, on **The Sandra Kublik Podcast**, Feb 2024
- Panel: *AI Safety Unplugged: Navigating the Risks Without the Hype* at **WEF, AI House Davos 2024**, with Yann LeCun, Max Tegmark, and David Haber
- *The Truth about AI: Lecture 1* at the **Royal Institution Christmas Lectures on BBC 4**, December 2023
- Panel on Misinformation in Generative AI at Widening NLP, **EMNLP 2024**
- *Fairness after shifts to transfer learning and generative NLP,* **University of Cambridge,** December 2023
- *Recent shifts in the field of Natural Language Processing and its implications for safety and fairness,* **University of Aberdeen,** Human-Centred AI Research Network AI Ethics and Safety Workshop, November 2023
- Royal Society Workshop on AI safety Risks, October 2023
- Panel at **ARIAS Amsterdam** on AI and the Arts, September 2023
- Panel talk on Responsible Generative AI with **Salesforce Ventures x Dawn Capital,** September 2023

- Responsible AI Webinar with **Five9**, 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**, January 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) and guest lectured 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
Twitter: https://x.com/seraphinagt