

**Arjun Subramonian** Email: [arjunsu@cs.ucla.edu](mailto:arjunsu@cs.ucla.edu)  
Website: [arjunsubramonian.github.io](https://arjunsubramonian.github.io)  
Publications: Semantic Scholar, Google Scholar  
GitHub: [github.com/ArjunSubramonian](https://github.com/ArjunSubramonian)

## Research Interests

AI, evaluation science, governance, alignment, fairness, social impacts

## Education

**PhD in Computer Science; University of California, Los Angeles (2025)**, GPA: 3.940  
Amazon Science Hub Fellow, NSF MENTOR Fellow, Eugene V. Cota-Robles Fellow  
*Advisors:* Yizhou Sun, Kai-Wei Chang

**BS in Computer Science; University of California, Los Angeles (2021)**, GPA: 3.927  
Summa Cum Laude

## Work Experience

Research Scientist; Meta FAIR AI & Society (2025)  
Research Intern; Meta FAIR Cost and Alignment (2024)  
Research Intern; Meta FAIR Society and Responsible AI (2022)  
Research Intern; Microsoft Research FATE (2022)  
Privacy Research Intern; Snap, Inc. (2021)  
AllenNLP Research Engineering Intern; Allen Institute for Artificial Intelligence (2021)  
Software Engineering Intern; Microsoft Corporation (2020)  
Software Engineering Intern; Get Heal, Inc. (2019)

## Conference Papers, Journal Papers, and Book Chapters

**Agree to Disagree? A Meta-Evaluation of LLM Misgendering,** [URL](#)  
**Arjun Subramonian**, Vagrant Gautam, Preethi Seshadri, Dietrich Klakow, Kai-Wei Chang, Yizhou Sun  
COLM 2025 (32% acceptance rate)

**An Effective Theory of Bias Amplification,** [URL](#)  
**Arjun Subramonian**, Samuel Bell, Levent Sagun, Elvis Dohmatob  
ICLR 2025 (32.08% acceptance rate)

**Strong Model Collapse,** [URL](#)  
Elvis Dohmatob, Yunzhen Feng, **Arjun Subramonian**, Julia Kempe  
ICLR 2025 **Spotlight** (5.1% acceptance rate)

**SHADES: Towards a Multilingual Assessment of Stereotypes in Large Language Models,** [URL](#)  
Margaret Mitchell, ..., **Arjun Subramonian**, ..., Zeerak Talat  
NAACL 2025 (22% acceptance rate)

**Theoretical and Empirical Insights into the Origins of Degree Bias in Graph Neural Networks,** [URL](#)  
**Arjun Subramonian**, Jian Kang, Yizhou Sun  
NeurIPS 2024 (25.8% acceptance rate)

**Understanding “Democratization” in NLP Research, [URL](#)**  
**Arjun Subramonian\***, Vagrant Gautam\*, Dietrich Klakow, Zeerak Talat  
EMNLP 2024 (20.8% acceptance rate)

**Networked Inequality: Preferential Attachment Bias in Graph Neural Network Link Prediction, [URL](#)**  
**Arjun Subramonian**, Levent Sagun, Yizhou Sun  
ICML 2024 (27.5% acceptance rate)  
Graph Learning Frontiers @ NeurIPS 2023 **Oral** (15.4% acceptance rate)

**Evaluating the Social Impact of Generative AI Systems in Systems and Society, [arXiv](#)**  
Irene Solaiman\*, Zeerak Talat\*, ..., **Arjun Subramonian**  
Oxford University Press Handbook on Generative AI

**Motif-Driven Contrastive Learning of Graph Representations, [URL](#)**  
Shichang Zhang, Ziniu Hu, **Arjun Subramonian**, Yizhou Sun  
Transactions on Knowledge and Data Engineering

**Factoring the Matrix of Domination: A Critical Review and Reimagination of Intersectionality in AI Fairness, [URL](#)**  
Anaelia Ovalle, **Arjun Subramonian**, Vagrant Gautam, Gilbert Gee, Kai-Wei Chang  
AIES 2023 **Oral** (7.7% acceptance rate)

**Bound by the Bounty: Collaboratively Shaping Evaluation Frameworks for Queer AI Harms, [URL](#)**  
Organizers Of QueerInAI, ..., **Arjun Subramonian**, ..., Jess de Jesus de Pinho Pinhal  
AIES 2023 (28.9% acceptance rate)

**It Takes Two to Tango: Navigating Conceptualizations of NLP Tasks and Measurements of Performance, [URL](#)**  
**Arjun Subramonian**, Xingdi Yuan, Hal Daumé III, Su Lin Blodgett  
Findings of ACL 2023 (41.89% acceptance rate)

**[Best Paper] Queer In AI: A Case Study in Community-Led Participatory AI, [URL](#)**  
Organizers of QueerInAI, ..., **Arjun Subramonian**, ..., Luke Stark  
FAccT 2023 (24.5% acceptance rate)

**On the Discrimination Risk of Mean Aggregation Feature Imputation in Graphs, [URL](#)**  
**Arjun Subramonian**, Kai-Wei Chang, Yizhou Sun  
NeurIPS 2022 (25.6% acceptance rate)

**Queer in AI, [URL](#)**  
Organizers of QueerInAI, ..., **Arjun Subramonian\***, ..., Jeffrey Xiong\*  
XRDS: Crossroads, The ACM Magazine for Students, Volume 28, Issue 4

**On Dyadic Fairness: Exploring and Mitigating Bias in Graph Connections, [URL](#)**  
**Arjun Subramonian**  
ICLR 2022 Blogpost Track (32% acceptance rate)

**Harms of Gender Exclusivity and Challenges in Non-Binary Representation in Language Technologies, [URL](#)**  
Sunipa Dev, Masoud MonajatiPoor\*, Anaelia Ovalle\*, **Arjun Subramonian\***, Jeff M Phillips, Kai-Wei Chang  
EMNLP 2021 Oral (9.8% acceptance rate)  
Women in Machine Learning @ NeurIPS 2021

**MOTIF-Driven Contrastive Learning of Graph Representations, [URL](#)**

## Arjun Subramonian

AAAI 2021 Undergraduate Consortium (17% acceptance rate)

### Automated, Cost-Effective Optical System for Accelerated Antimicrobial Susceptibility Testing (AST) Using Deep Learning, [URL](#)

Calvin Brown, ..., Arjun Subramonian, ..., Aydogan Ozcan

ACS Photonics

### Estimating the Ages of FGK Dwarf Stars Through the Use of GALEX FUV Magnitudes, [URL](#)

Sara Crandall, Graeme H. Smith, Arjun Subramonian, Kelly Ho, Evelyn M. Cochrane

The Astronomical Journal

## Workshop Papers

### Fairness Implications of GNN-to-MLP Knowledge Distillation, [URL](#)

Margaret Capetz, Yizhou Sun, Arjun Subramonian

Reliable ML @ NeurIPS 2025

### Challenges to Grassroots Organization Engagement with AI Policy, [URL](#)

Jennifer Mickel, ..., Arjun Subramonian

Algorithmic Collective Action @ NeurIPS 2025 [Oral](#)

### Issues in Measuring the Fairness of Social Representation in Synthetic (Speech) Data, [URL](#)

Arjun Subramonian, Brooklyn Sheppard, Levent Sagun

Synthetic Data @ Aarhus Conference 2025

### Pairwise Matching of Intermediate Representations for Fine-grained Explainability, [arXiv](#)

Lauren Shrack, Timm Haucke, Antoine Salaün, Arjun Subramonian, Sara Beery

CV4Animals @ CVPR 2025 [Oral](#)

### Stop! In the Name of Flaws:

#### Disentangling Personal Names and Sociodemographic Attributes in NLP, [URL](#)

Vagrant Gautam, Arjun Subramonian, Anne Lauscher, Os Keyes

Gender Bias in NLP @ ACL 2024

### Prompting Multilingual Large Language Models to Generate Code-Mixed Texts:

#### The Case of South East Asian Languages, [URL](#)

Zheng-Xin Yong, ..., Arjun Subramonian, ..., Alham Fikri Aji

Computational Approaches to Linguistic Code-Switching @ EMNLP 2023

### Critical Technopolicy and Reflexivity For Proactive Gender-Inclusive NLP

Davi Liang, Anaelia Ovalle, Arjun Subramonian, Alicia Boyd

Queer in AI @ ACL 2023

### Group Excess Risk Bound of Overparameterized Linear Regression with Constant-Stepsize SGD, [URL](#)

Arjun Subramonian, Levent Sagun, Kai-Wei Chang, Yizhou Sun

Trustworthy and Socially Responsible Machine Learning @ NeurIPS 2022

### You Reap What You Sow: On the Challenges of Bias Evaluation Under Multilingual Settings, [URL](#)

Zeerak Talat, ..., Arjun Subramonian\*, ..., Oskar van der Wal\*

Challenges & Perspectives in Creating Large Language Models @ ACL 2022

**How to Make Virtual Conferences Queer-Friendly: A Guide**, [URL](#)

Organizers of QueerInAI, ..., **Arjun Subramonian**, ..., Nyx McLean

Widening NLP @ EMNLP 2021

**Motif-Driven Contrastive Learning of Graph Representations**, [arXiv](#)

Shichang Zhang, Ziniu Hu, **Arjun Subramonian**, Yizhou Sun

Self-Supervised Learning @ WWW 2021

## Preprints

**OpenApps: Simulating Environment Variations to Measure UI Agent Reliability**

Karen Ullrich, ..., **Arjun Subramonian**, ..., Mark Ibrahim

**auto-fpt: Automating Free Probability Theory Calculations for Machine Learning Theory**, [arXiv](#)

**Arjun Subramonian**, Elvis Dohmatob

**Survey of Bias In Text-to-Image Generation: Definition, Evaluation, and Mitigation**, [arXiv](#)

Yixin Wan, **Arjun Subramonian**, ..., Kai-Wei Chang

**Weisfeiler and Leman Go Measurement Modeling: Probing the Validity of the WL Test**, [arXiv](#)

**Arjun Subramonian**, Adina Williams, Maximilian Nickel, Yizhou Sun, Levent Sagun

**BLOOM: A 176B-Parameter Open-Access Multilingual Language Model**, *in submission*, [arXiv](#)

Teven Le Scao, ..., **Arjun Subramonian**, ..., Thomas Wolf

**Fairness and Bias Mitigation: A practical guide into the AllenNLP Fairness module**, [URL](#)

**Arjun Subramonian**

## Public Scholarship

**Queer AI Harms and Proposed Policy Interventions**, [URL](#)

Organizers of QueerInAI, ..., **Arjun Subramonian**, ...

**AI and Queer Communities**, [URL](#)

**Arjun Subramonian**

**FINDINGS: Exploring the Impact of AI**, [URL](#)

National AI Advisory Committee, ..., **Arjun Subramonian**, ...

**Queer in AI National AI Advisory Committee 2023 Briefing**, [URL](#)

Organizers of QueerInAI, **Arjun Subramonian**

**Rebuilding Trust: Queer in AI Approach to Artificial Intelligence Risk Management**, [URL](#)

Organizers of QueerInAI, Ashwin S\*, William Agnew\*, Hetvi Jethwani\*, **Arjun Subramonian\***

## Invited Talks

2025 - Guest Lecture: A Critical Review and Reimagination of Intersectionality and Democratization in AI,

Non-Ideal Algorithmic Fairness, Cornell Tech

2025 - An Effective Theory of Bias Amplification, ACM AI at UCLA Reading Group

2025 - Agree to Disagree? A Meta-Evaluation of LLM Misgendering, Meta FAIR; [Slides](#)

2025 - An Effective Theory of Bias Amplification, USC ISI Natural Language Seminar; [Event](#), [Slides](#), [Video](#)

2025 - Theoretical and Empirical Insights into Degree Bias in GNN, DEFIRST Reading Group - MILA x Vector; [Video](#)

2025 - From “Democratization” to Personal Names: Reimagining NLP Practices Towards Justice, USC ISI Natural Language Seminar; [Event](#), [Slides](#), [Video](#)

2024 - From “Democratization” to Personal Names: Reimagining NLP Practices Towards Justice, Algorithmic Decision Systems (ADS) Reading Group, UC Berkeley; [Slides](#)

2024 - Queer In AI: A Case Study in Community-Led Participatory AI, Microsoft GLEAM Security; [Slides](#)

2024 - From “Democratization” to Personal Names: Reimagining NLP Practices Towards Justice, UC Berkeley Data Science for Social Justice Workshop; [Slides](#)

2024 - ‘Queer-Inclusive’ AI: Contesting the Power Structures and Logics of Machine Learning, IBM Research Tech for Justice Seminar; [Slides](#)

2024 - Queer In AI: A Case Study in Community-Led Participatory AI, Data & Society; [Slides](#)

2023 - Bias, Discrimination, and Power: Graphs and Natural Language, KUNGFU.AI; [Slides](#)

2023 - Bias and Power in NLP, USC ISI Natural Language Seminar; [Event](#), [Slides](#), [Video](#)

2022 - Bias and Power in NLP (for future consultants), Paris; [Slides](#)

2022 - Bias and Power in NLP, NLP Seminars at Dublin College University; [Slides](#)

2022 - Guest Lecture: Bias in Natural Language Processing, COM SCI 263: NLP, UCLA; [Slides](#), [Video](#)

2022 - Queer in AI: Making AI Queer-Inclusive and Prioritizing Grassroots D&I Activism, Humlab, Umeå University; [Slides](#), [Video](#)

2022 - Prioritizing Grassroots D&I Activism: Queer in AI, AAAI Workshop on Diversity in Artificial Intelligence; [Event](#), [Slides](#)

2022 - Prioritizing Grassroots D&I Activism: Queer in AI, Nike Sport+AI Conference; [Event](#), [Slides](#)

2021 - How Can I Make My Hackathon Queer-Inclusive?, Hackcon IX; [Slides](#), [Video](#)

2019 - An Automated and Cost-Effective System for Early Antimicrobial Susceptibility Testing Using Optical Fibers and Deep Learning, UCLA HHMI Day; [Slides](#)

## Invited Panels

2025 - Asian Queer Experiences in Engineering: Intersectionality, Inclusion, and Institutional Change, American Society for Engineering Education

2024 - QICSE Panel, UC Berkeley

ICML Queer in AI Workshop; [Event](#)

2024 - Challenges and Perspectives for queer and/or disabled communities presented by AI in HCI, ICML Queer in AI Workshop; [Event](#)

2024 - Artificial Intelligence and Democratic Values Roundtable, Center for AI & Digital Policy; [Event](#)

2024 - UCLA Computer Science PhD Visit Day Panel

2024 - UCLA ACM AI Outreach GALA Panel

2024 - Queer Computation: A Critical Unmaking, Coding Inquiry, McMaster

2024 - AI in the Time of Generative Models: Ensuring that Diverse Voices Are Heard, AAAI Workshop on Diversity in AI; [Event](#)

2023 - Students with Disabilities Panel, UCLA

2023 - US AI Policy: Next Steps

2023 - Queer in AI National AI Advisory Committee Briefing; [Event](#), [Briefing](#), [Video](#), [Press](#), [Findings](#)

2023 - Partnership on AI Inclusive Data Collection Workshop

2023 - Data & Society Salon Series on Responsible AI in Government

2022 - Students with Disabilities Panel, UCLA

2022 - Accessibility and Inclusion Panel, Neuromatch Academy; [Video](#)

2022 - Gender as a Variable in NLP, NAACL Queer in AI Workshop; [Event](#), [Video](#)

2022 - NYU Center for Responsible AI Roundtable on Accessible AI Education

2022 - Co-Opting AI: Queer, NYU’s Institute for Public Knowledge; [Event](#), [Video](#)

2022 - How Do We Improve DEI in AI?, Nike Sport+AI Conference  
2021 - Eye on A.I.: Equity & Inclusion in A.I. Technology, Toronto Public Library; [Video](#)  
2021 - Intersectionality Panel, NAACL; [Video](#)

## Honors

2024 - Two Sigma PhD Fellowship Nominee, 1 of 6 students selected by UCLA  
2024 - Amazon Science Hub Fellowship; [Site](#)  
2024 - Google PhD Fellowship Nominee, 1 of 4 students selected by UCLA  
2024 - JP Morgan PhD Fellowship Nominee, 1 of 2 students selected by UCLA  
2023 - LoG Conference Top-10 Reviewer  
2023 - Queer in AI Ford Foundation Research Grant; [Site](#)  
2023 - FAccT Best Paper Award; [Site](#)  
2022 - NSF MENTOR Fellowship  
2021 - Allen AI Outstanding Intern of the Year; [Site](#)  
2021 - Major League Hacking Top 50; [Site](#)  
2021 - UCLA Samueli School-Wide Outstanding Bachelor of Science; [Site](#)  
2021 - UCLA Chancellor's Service Award; [Site](#)  
2021 - UCLA Samueli Engineering Achievement Award in Student Welfare; [Site](#)  
2021 - UCLA Eugene V. Cota-Robles Fellowship  
2021 - Boeing Company Scholarship  
2021 - Brian J. Lewis Endowment  
2018-2021 - UCLA Dean's Honors List  
2020 - Computing Research Association Outstanding Undergraduate Researcher Honorable Mention; [Site](#)  
2020 - IBM Quantum Challenge  
2019 - 3rd Place Award for Best Hack @ Rose Hack, Major League Hacking  
2017 - Siemens Competition Regional Finalist; [Site](#)  
2016 - Award of Achievement, Association for Computing Machinery, San Francisco Bay Area Professional Chapter

## Service

- **Reviewing:**
  - **Conferences:**
    - \* NeurIPS (2024, D&B 2024, Ethics 2024, 2023, D&B 2023, Ethics 2023)
    - \* ICML (2025)
    - \* ICLR (2025, Workshops 2025)
    - \* FAccT (2025, 2024, 2023, 2022)
    - \* LoG (2024, 2023, 2022)
    - \* KDD (2024, 2023)
    - \* ACL (Ethics 2024)
    - \* IASEAI (2026)
    - \* WACV (2026)
    - \* ICCV (2025)
    - \* WebConf (2024)
  - **Workshops:**
    - \* NeurIPS (GLFrontiers 2023, GLFrontiers 2022, TSRML 2022)
    - \* NAACL (TrustNLP 2022, WOAH 2022, SRW 2022)
    - \* ACL (TrustNLP 2023, WOAH 2023, Queer in AI 2023, BigScience 2022)

- Conference organization:
  - LoG Outreach Chair ([2024](#))
  - FAccT Socials Chair ([2024](#))
  - NeurIPS Affinity Workshops Chair ([2022](#), [2023](#))
  - KDD EDI Day Chair ([2023](#))
  - NAACL Accessibility Chair ([2022](#))
- Workshop and social organization:
  - [Evaluating Evaluations: Examining Best Practices for Measuring Broader Impacts of Generative AI @ NeurIPS 2024](#)
  - In Código of Hope: Possibilities of Global Solidarity and Activism in Academic Environments @ FAccT 2024
  - [Global AI Cultures @ ICLR 2024](#)
  - [Collaboratively Developing Evaluation Frameworks for Queer AI Harms @ FAccT 2022](#)
  - [Queer in AI: ICML \(2024, 2022, 2021\); NeurIPS \(2021\); ICLR \(2024\); AAAI \(2021\); NAACL \(2025, 2022\); ACL \(2023\)](#)
- Policy:
  - [Queer in AI Policy Working Group](#)
  - [NIST AI Safety Institute Consortium](#)
- Teaching:
  - [UCLA: Computer Science 32 \(2023\), Computer Science Summer Institute \(2022, 2023\)](#)

## Mentoring

I am fortunate to have mentored brilliant students:

- Margaret Capetz (PhD at University of Washington)
- Naisha Agarwal
- Pranav Subbaraman
- Steven Swee (PhD at UCLA)

## References

Yizhou Sun ([yzsun@cs.ucla.edu](mailto:yzsun@cs.ucla.edu))  
 Kai-Wei Chang ([kwchang@cs.ucla.edu](mailto:kwchang@cs.ucla.edu))  
 Levent Sagun ([leventsagun@meta.com](mailto:leventsagun@meta.com))  
 Elvis Dohmatob ([e.dohmatob@meta.com](mailto:e.dohmatob@meta.com))  
 Max Nickel ([maxn@meta.com](mailto:maxn@meta.com))