Musashi Hinck

Princeton, NJ musashi.hinck@gmail.com (609) 786-0299 drmusashihinck
muhark

About Me

AI Research Scientist studying how to characterize, measure, and correct latent social biases in generative models.

Education

University of Oxford, *Doctor of Philosophy* — *Oxford, UK*
Thesis on applications of ML/DL to political campaigns: — *Oct 2019 – Aug 2022*
- Demonstrating the effectiveness of offline RL for targeting political messages.
- Using LLMs for sample selection in large newspaper corpora.
- Investigating the suitability of explainable AI tools for statistical inference.

University of Oxford, *Master of Science* — *Oxford, UK*
Thesis used unsupervised NLP to study parliamentarians' use of Twitter. — *Oct 2018 – Sept 2019*

University of Oxford, *Bachelor of Arts* — *Oxford, UK*
Read Philosophy, Politics and Economics at Merton College. — *Oct 2012 – July 2015*

Experience

AI Research Scientist — *USA (Remote)*
Intel Labs — *Feb 2024 – present*
AI Research Scientist in Multimodal Cognitive AI group, Emergent AI Lab
- Received Intel Labs Scholar Award Q3'24 for highest publication output (10 papers in first year).
- Executed large multimodal model training to over 1k accelerators.
- Developed highly optimized workloads on cutting-edge AI hardware.

Postdoctoral Research Associate — *Princeton, NJ, USA*
Princeton University — *Oct 2022 – Feb 2024*
Researcher at Data-Driven Social Science Initiative working with Professor Brandon M. Stewart on:
- Statistical methods for removing bias from analyses relying on LLM predictions.
- Evaluating political influence of state propaganda in LLM training data.
- Innovating methods for controlled text generation for causal inference.

Predoctoral Research Fellow — *London, UK*
University College London — *2021 – 2022*
Developed deep learning (DL) and natural language processing (NLP) methods for £1M UKRI-funded research project on Mental Models of the Political Economy.

Data Scientist — *London, UK*
Kanto Systems Ltd. — *2019 – 2022*
Used machine learning to develop voter profiles from survey and geographic data.

Quantitative Researcher — *Tokyo, Japan*
SBI Holdings Ltd. — *2017 – 2018*
Provided business intelligence and predictive analytics for clients and internal partners at Japan's largest private stock exchange, SBI Japannext.

Public Working Papers

Semantic Specialization in Moe Appears with Scale: A Study of DeepSeek-R1 Expert Specialization
Matthew Lyle Olson*, Neale Ratzlaff*, Musashi Hinck*, Man Luo, Sungduk Yu, Chendi Xue, Vasudev Lal.

IssueBench: Millions of Realistic Prompts for Measuring Issue Bias in LLM Writing Assistance
Paul Röttger, Musashi Hinck, Valentin Hoffman, Kobi Hackenburg, Valentina Pyatkin, Faeze Brahman, Dirk Hovy.

Publications

Debiasing Large Vision-Language Models by Ablating Protected Attribute Representations
Neale Ratzlaff, Matthew Lyle Olson, Musashi Hinck, Shaoyen Tseng, Vasudev Lal, Phillip Howard. *NeurIPS 2024 SafeGenAI Workshop*

Exploring Vision Transformers for Early Detection of Climate Change Signals
Sungduk Yu, Anahita Bhiwandiwall, Yaniv Gurwicz, Musashi Hinck, Matthew Lyle Olson, Raanan Rohekar, Vasudev Lal. *NeurIPS 2024 Tackling Climate Change with Machine Learning*

AutoPersuade: A Framework for Evaluating and Explaining Persuasive Arguments
Till Raphael Saenger, Musashi Hinck, Justin Grimmer, Brandon M. Stewart. *EMNLP 2024 (Main)*

Why do LLaVA Vision-Language Models Reply to Images in English?
Musashi Hinck, Carolin Holtermann, Matthew Lyle Olson, Florian Schneider, Sungduk Yu, Anahita Bhiwandiwalla, Anne Lauscher, Shaoyen Tseng, Vasudev lal. *EMNLP 2024 (Findings)*

LLaVA-Gemma: Accelerating Multimodal Foundation Models with a Compact Language Model
Musashi Hinck*, Matthew Lyle Olson*, David J. Cobbley, Shaoyen Tseng, Vasudev Lal. *CVPR 2024 Multimodal Foundation Models Workshop*

Political Compass or Spinning Arrow? Towards More Meaningful Evaluations for Values and Opinions in Large Language Models
Paul Röttger, Valentin Hoffman, Valentina Pyatkin, Musashi Hinck, Hannah Rose Kirk, Hinrich Schütze, Dirk Hovy. *ACL 2024*

Using Imperfect Surrogates for Downstream Inference: Design-based Supervised Learning for Social Science Applications of Large Language Models
Naoki Egami, Musashi Hinck, Hanying Wei, Brandon M. Stewart. *NeurIPS 2023*

Software, Models and Data

IssueBench *2025*
A realistic benchmark of 2.49M prompts for evaluating issue bias in LLMs.
- In collaboration with Paul Röttger, Valentin Hoffman, Kobi Hackenburg, Valentina Pyatkin, Faeze Brahman and Dirk Hovy.

LLaVA-Gemma *2024*
A small vision-language model finetuned from Google's Gemma models.
- Start-to-finish project in first month at Intel Labs
- In collaboration with Matthew Lyle Olson, David J. Cobbley, Shaoyen Tseng and Vasudev Lal.

Design-based Supervised Learning *2024*
A statistical framework for getting statistically valid estimates from LLM predictions.
- Based on framework proposed in Hinck, Egami, Wei and Stewart (NeurIPS 2023).
- In collaboration with Naoki Egami, Hanying Wei and Brandon M. Stewart.

Skills

Programming Languages: Experienced with Python, R and Shell

AI Engineering: Experienced with PyTorch, HuggingFace, deepspeed

Dev/MLOps: Experienced with k8s and SLURM for cluster management, wandb for experiment tracking

(Natural) Languages: English (native), Japanese (proficient/native), Spanish (proficient, CEFR C1)