Mehar Bhatia

Ph.D. Student in Computer Science | McGill University & MILA

mehar.bhatia@mila.quebec | Montreal, CA | meharbhatia.github.io | meharbhatia | bhatia_mehar | Scholar

Research Focus

I am a second-year PhD student at MILA and McGill University, advised by Dr. Siva Reddy and Dr. Vered Shwartz. My research is centered around **socio-technical alignment** of AI systems to represent **diverse human preferences** and **maintain diversity**. I investigate how language and multimodal models can learn from real-world human interaction, through adaptive optimization and personalized reasoning. My work aims to model and reduce epistemic uncertainty in human-AI collaboration, enabling systems capable of continual learning, and mutual adaptation.

Keywords: AI Alignment, LLM Post-training, LLM Personalization, AI Safety, Continual Learning, Interpretability

Education

Sept 2024 - Present | **McGill University | MILA Quebec AI Institute** | **Montreal, Canada**
Doctor of Philosophy (Ph.D.) in Computer Science, *GPA: 4.0*
Advisors: Dr. Siva Reddy & Dr. Vered Shwartz (UBC)

Sept 2022 - Aug 2024 | **University of British Columbia | Vector Institute** | **Vancouver, Canada**
Master of Science (MS) in Computer Science, *GPA: 4.0*
Advisor: Dr. Vered Shwartz

July 2016 - June 2020 | **Shiv Nadar University** | **NCR, India**
B.Tech in Computer Science and Engineering, Minor in Mathematics

Experience

Sept 2024 - Present | **McGill NLP Lab [🌐] & MILA Quebec AI Institute [🌐]** | **Montreal, Canada**
PhD Student and Researcher with Prof. Siva Reddy & Prof. Vered Shwartz
- Actively pursuing several research projects, across LLM alignment, AI safety, societal impacts of AI, multimodality, reasoning and personalization [see works under Relevant Publications].

Jan 2023 - Aug 2024 | **UBC NLP Lab [🌐] and Vector Institute of AI [🌐]** | **Vancouver, Canada**
MSc Student and Researcher with Prof. Vered Shwartz
- Developed **Geo-Diverse COMET** model for generating culturally relevant commonsense inferences.
- Created a multicultural image-text **CulturalSnap** dataset (∼75K images) for analyzing cultural understanding in vision-language (VL) models using contrastive learning and designed **two benchmarks** for cultural retrieval and visual grounding, revealing gaps in VL models' cultural awareness.
- In collaboration with UWashington & AI2, developed large-scale **CulturalBench** to evaluate and track LLMs' cultural knowledge and performance across varied cultural contexts.

July 2021 - July 2022 | **NeuralSpace [🌐]** | **Remote / London, UK**
Applied Research Scientist | Mentors: Felix Laumann (CEO) and Ayushman Dash (CTO)
- Worked on understanding the disparity between research and deployment for low-resource languages and developing technologies to mitigate this gap (*'Language Technologies for All'*).
- Incorporated pipelines for joint multiple intent detection and slot filling using contrastive learning. Benchmarked accuracy improvement of **30%** when compared to HF models for Indic languages.

June 2020 - June 2021 | **IIIT Delhi | Multimodal Digital Media Analysis Lab [🌐]** | **New Delhi, India**
Research Assistant | Advisor: Prof. Rajiv Ratn Shah
- Devised approaches for attributing the prediction of neural AES models to its input features.
- Responsible for developing ASR for Japanese-accented speech using self-supervised methods. Deployed and used by **300,000** middle-school students.
- Identified cognitive theories on bilingual language acquisition using visual lip-reading models.
- *Interpretability* *Speech Processing* *Vision*

Jan 2020 | Research Intern | Advisors: Prof. Rajiv Ratn Shah and Prof. Junyi Jessi Li
- Proposed a model-agnostic adversarial suite to evaluate the robustness of Automatic Essay Scoring (AES) systems, and presented associated metrics to test natural language understanding capabilities.
- Implemented a pipeline to identify bias, disparate error rates with respect to different groups for AES.
- *Question Answering* *Adversarial Robustness* *NLU* *Fairness*

June 2019 - Aug 2019 | **IIIT Delhi | Multimodal Digital Media Analysis Lab** [🌐] **New Delhi, India**
Research Intern | Advisors: Prof. Rajiv Ratn Shah and Dr. Debanjan Mahata
- Proposed a novel solution for Automatic Knowledge Graph Construction from unstructured text across multiple domains. **Awarded Honorable Mention Prize** and **Travel Grant** at ICDM 2019 in Beijing, China.
- *Knowledge Graphs* *Event Coreference Resolution*

May 2019 - July 2019 | **Languages Technologies Research Centre | NLP and MT Lab, IIIT-H** [🌐] **Hyderabad, India**
Summer Research Intern | Advisors: Prof. Dipti Misra Sharma and Prof. Manish Shrivastava
- Identified linguistic factors crucial for the performance of Statistical and Neural Machine Translations and implemented seminal papers in Neural MT for English – Hindi translations.
- *Neural Machine Translation* *Indic Languages*

May 2018 - July 2018 | **Software Engineering Research Centre | IIIT-H** [🌐] **Hyderabad, India**
Summer Research Intern | Advisor: Prof. Y Raghu Reddy
- Surveyed, implemented and analyzed the performance of state-of-the-art algorithms and developed an end-to-end semi-automated ontology enrichment pipeline using a sequential deep learning model.
- *Ontologies* *NLP* *Machine Learning*

Relevant Publications

S=In Submission, C=Conference, W=Workshop, P=Preprint

[S.2] **Value Drifts: Tracing Value Alignment During LLM Post-Training** [🔗]
Mehar Bhatia, Shravan Nayak, Gaurav Kamath, Marius Mosbach, Karolina Stanczak, Vered Shwartz, and Siva Reddy
The Fourteenth International Conference on Learning Representations [**Under Review**]

[S.1] **DeepSeek-R1 Thoughtology: Let's think about LLM reasoning** [🔗]
Sara Vera Marjanovic, Arkil Patel, Adlakha, Aghajohari, BehnamGhader, Mehar Bhatia, Aditi Khandelwal,..., Amirhossein Kazemnejad, Gaurav Kamath, Marius Mosbach, Karolina Stanczak, Siva Reddy
Transactions on Machine Learning Research (TMLR) [**Under Review**]

[W.1] **Societal Alignment Frameworks Can Improve LLM Alignment** [🔗]
Karolina Stanczak, Nicholas Maede, Mehar Bhatia, Hattie Zhou,..., Sylvie Delacroix, Gillian K. Hadfield, Siva Reddy
Bidirectional Human-AI Alignment Workshop at ICLR'25 and presented at IASEAI '25 [**BiAlign @ ICLR'25**]

[C.6] **Assessing Cultural Expectation Alignment in Text-to-Image Models and Evaluation Metrics** [🔗]
Shravan Nayak, Mehar Bhatia, Xiaofeng Zhang, Verena Rieser, Lisa Anne Hendricks, Sjoerd van Steenkiste, Yash Goyal, Karolina Stanczak, Aishwarya Agrawal
The 2025 Conference on Empirical Methods in Natural Language Processing [**EMNLP'25 Findings**]

[C.5] **From Local Concepts to Universals: Evaluating Multicultural Understanding of Vision-Language Models** [🔗]
Mehar Bhatia, Sahithya Ravi, Aditya Chinchure, Eunjeong Hwang, Vered Shwartz
The 2024 Conference on Empirical Methods in Natural Language Processing [**EMNLP'24, Main**]

[C.4] **CulturalTeaming: AI-Assisted Interactive Red-Teaming for Challenging LLMs' Multicultural Knowledge** [🔗]
Yu Ying Chiu, Liwei Jiang, Maria Antoniak, Chan Young Park, Shuyue Stella Li, Mehar Bhatia, Sahithya Ravi, Yulia Tsvetkov, Vered Shwartz, Yejin Choi
The 63rd Annual Meeting of the Association for Computational Linguistics [**ACL'25, Main**]

[C.3] **GD-COMET: A Geo-Diverse Commonsense Inference Model** [🔗]
Mehar Bhatia, Vered Shwartz
The 2023 Conference on Empirical Methods in Natural Language Processing [**EMNLP'23, Main**]

[C.2] **One To Rule Them All: Towards Joint Indic Language Hate Speech Detection** [🔗]
Mehar Bhatia, Tenzin Singhay Bhotia, Akshat Agarwal, Kumar Shridhar, Felix Laumann, Ayushman Dash
Forum for Information Retrieval Evaluation [**FIRE'21**]

[C.1] **Evaluation Toolkit For Robustness Testing Of Automatic Essay Scoring Systems** [🔗]
Anubha Kabra*, Mehar Bhatia*, Yaman Kumar Singla*, Junyi Jessy Li, Di Jin, Rajiv Ratn Shah (* = Equal Contribution)
ACM International Conference on Data Science & Management of Data [**CODS-COMAD'21**]

Honours and Awards

Awarded FRQNT Doctoral Training Research Scholarship [🔗] (amount $100,000) Montreal, Quebec, April 2025
Awarded MILA EDI Scholarship - Women in AI Category [🔗] (amount $24,000) MILA, Montreal, March 2025
Awarded Prof. Cho Diversity Scholarship (amount $1,500) MILA, Montreal, Dec 2024
Awarded Grad Excellence Award in Computer Science (amount $6,000) McGill University, Montreal, Sept 2024
Awarded BPOC Graduate Excellence Award (amount $1,650) University of British Columbia, Vancouver, April 2024
Awarded Vector Research Grant (amount $8,000) Vector Institute for Artificial Intelligence, Toronto, 2023 & 2024
Granted International Tuition Award (amount $6,400) University of British Columbia, Vancouver Sept 2022
Awarded for Top 5 BTech Thesis Projects in Computer Science Department Shiv Nadar University, 2020
Complete Tuition Fee Wavier for Undergraduate Program Shiv Nadar University (2016-2020)

Media

La Presse, Canada (April 2024) Imagine the future of AI on the beach [Article: 🌐]
Global News, Canada (Jan 2024) Solving Diversity Issues in Artificial Intelligence Models [Live TV Interview: 🌐]
UBC News (Jan 2024) ChatGPT has read almost the whole internet. That hasn't solved its diversity issues [Article: 🌐]
Radio-Canada (Jan 2024) Même après avoir lu presque tout le web, l'intelligence artificielle a des préjugés [Article: 🌐]
L'Obs (Jan 2024) ChatGPT, une intelligence artificielle pleine de biais (mais ça se soigne) [Article: 🌐]
The Conversation Canada (Feb 2024) AI needs to be trained on culturally diverse datasets to avoid bias) [Article: 🌐]

Technical Skills

Programming Languages Python Java R C JavaScript SQL HTML/CSS
Technologies PyTorch Tensorflow Keras NLTK MySQL MongoDB LaTex Git

Talks

UBC Computer Science (Nov 2025) Value Drifts: Tracing Value Alignment During LLM Post-Training [Slides: 🌐]
MILA AI-Safety RG (Nov 2025) Persona Features Control Emergent Misalignment [Slides: 🌐]
UBC CS MSc Presentation (Aug 2024) Exploring Cultural Competence in Language and Multimodal Models [Slides: 🌐]
Guest Lecturer at UBC's CPSC 532V (Grad) (Feb 2024) Exploring Cross-Cultural Phenomena in NLP [Slides: 🌐]

Teaching Experience

Teaching Assistant (Sept-Dec 2025) McGill COMP 545 - Natural Language Understanding with DL (Prof. Siva Reddy)[🌐]
Teaching Assistant (Jan-April 2025) McGill COMP 345 - From Natural Language to Data Science (Prof. Siva Reddy)[🌐]
Teaching Assistant (April 2023, 2024) Prompt Engineering Labs conducted by Vector Institute with 120 participants
Graduate Teaching Assistant (Sept-Dec 2022) Course UBC CPSC 436N - NLP by Prof. Vered Shwartz [🌐]

Service and Workshops

Organizing Committee CVPR 2025 - VLMs4All Workshop
Reviewer ICLR'26 COLM'25 ACL'25 NAACL'25 EMNLP'24 AAAI'24
Summer School Participated in CIFAR Deep Learning & Reinforcement Learning (DLRL) Summer School 2023
Invitational Workshops (1) Safety-Guaranteed LLMs Workshop - Special Year on Large Language Models and Transformers, held at Simons Institute, UC Berkeley, April 2025
(2) Bellairs Invitational Workshop on Contemporary, Foreseeable and Catastrophic Risks of Large Language Models held in Barbados, April 2024