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

My research is centered around **socio-technical alignment** of AI systems to reflect **diverse human preferences** and **preserve diversity**. I investigate how language and multimodal models can continually learn from real-world human interactions through dynamic preference elicitation and adaptive post-training for efficient, personalized alignment.
Keywords: *AI Alignment*, *LLM Post-training*, *LLM Personalization*, *AI Safety*, *Continual Learning*, *Interpretability*

Education

Sept 2024 - 2028 (Exp.) — **McGill University | MILA Quebec AI Institute** — **Montreal, Canada**
Doctor of Philosophy (Ph.D.) in Computer Science, *GPA: 4.0*
Advisors: Prof. Siva Reddy & Prof. Vered Shwartz (UBC)
Fellowships: *FRQNT Doctoral Training Research Scholarship*

Sept 2022 - Aug 2024 — **University of British Columbia | Vector Institute** — **Vancouver, Canada**
Master of Science (MSc Research) in Computer Science, *GPA: 4.0*
Advisor: Prof. Vered Shwartz
Thesis: *Exploring Cultural Competence in Language and Multimodal Models* [🔗]

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**
Graduate Researcher
- Actively pursuing several research projects, across LLM alignment **[P10, P9]**, AI safety, societal impacts of AI, multimodality **[P7]**, reasoning **[P8]** and personalization *[see works under Relevant Publications]*.
- Investigated post-training dynamics in LLM value alignment, analyzing how training algorithms (SFT and preference optimisation, i.e., RLHF and variants) and datasets shape model behaviour **[ref: P10]**.

Jan 2023 - Aug 2024 — **UBC NLP Lab [🌐] and Vector Institute of AI [🌐]** — **Vancouver, Canada**
Graduate Researcher
- Developed 'GD-COMET' model for generating culturally relevant commonsense inferences **[ref: P4]**.
- Created a large-scale, image-text dataset across 50 diverse cultures, and designed training objectives for inclusive representation learning. Also designed 'two benchmarks' for cultural retrieval and visual grounding, revealing gaps in multicultural understanding of vision-language models **[ref: P6]**.
- In collaboration with AI2, developed a large-scale benchmark CulturalBench to evaluate and track LLMs' cultural knowledge and performance across varied cultural contexts **[ref: P5]**.

July 2021 - July 2022 — **NeuralSpace [🌐]** — **Remote / London, UK**
Applied Research Scientist
- 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 27% when compared to HF models for Indic languages **[ref: P3]**.
- Expanded SLU pipelines to end-to-end transformer-based architectures for Indic and African 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 by Benesse Corporation (Japanese education firm) and used by over 300,000 middle-school students.
- Identified cognitive theories on bilingual language acquisition using visual lip-reading models.

Jan 2020 -May 2020 | **IIIT Delhi | Multimodal Digital Media Analysis Lab** [🌐] **New Delhi, India**
Research Intern | Advisors: Prof. Rajiv Ratn Shah and Prof. Junyi Jessi Li (UT Austin)
- 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 **[ref: P2]**.
- Implemented a pipeline to identify bias, disparate error rates with respect to different groups for AES.

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.

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.

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 various algorithms and developed an end-to-end semi-automated ontology enrichment pipeline using a sequential deep learning model **[ref: P1]**.

Relevant Publications

[P10] **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, ICLR 2026 [**Under Review**]

[P9] **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**]

[P8] **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**]

[P7] **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**]

[P6] **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**]

[P5] **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**]

[P4] **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**]

[P3] **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**]

[P2] **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**]

[P1] **A survey on Ontology Enrichment from Text** [🔗]
Vivek Iyer, Lalit Mohan, Mehar Bhatia, Y Raghu Reddy
16th International Conference on Natural Language Processing, Hyderabad, India [**ICON'19**]

Honours and Awards

Fonds de recherche du Québec – Nature et technologies (FRQNT) Doctoral Training Research Scholarship [🔗] Montreal, Quebec, April 2025

MILA EDI Scholarship - Women in AI Category [🔗] MILA, Montreal, March 2025

Prof. Cho Diversity Scholarship MILA, Montreal, Dec 2024

Grad Excellence Award in Computer Science McGill University, Montreal, Sept 2024

BPOC Graduate Excellence Award University of British Columbia, Vancouver, April 2024

Vector Research Grant Vector Institute for Artificial Intelligence, Toronto, 2023 & 2024

International Tuition Award University of British Columbia, Vancouver Sept 2022

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)

Leadership and Service

Reviewer	ICLR'26 COLM'25 ACL'25 NAACL'25 EMNLP'24 AAAI'24
Organizing Committee	(1) CVPR 2026 - Multimodal Alignment for a Pluralistic Society (MAPS) Workshop (2) CVPR 2025 - VLMs4All Workshop
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

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 [🌐]

Talks

UBC Computer Science (Nov 2025) Value Drifts: Tracing Value Alignment During LLM Post-Training [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: 🌐]

Technical Skills

Programming Languages	Python Java R C JavaScript SQL HTML/CSS
Technologies	PyTorch Transformers HuggingFace vllm Tensorflow Keras MongoDB LaTex Git

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: 🌐]