Mehar Bhatia

Ph.D. Candidate 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 **shaping and sustaining aligned model behaviour**: **robustly** through character training, and **adaptively** through continual alignment. I study the values and personas models acquire during training, how these generalize, and how models can efficiently adapt to evolving users and environments while managing memory and preserving stability and privacy. I translate observed bottlenecks into evals, data, reward signals for model improvement.
Keywords: *AI Alignment*, *Character Training*, *Continual Alignment*, *Reward Modeling*, *Post-training*, *Behavioural Evaluation*

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* [🔗] (*Government of Québec*)

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 underlying datasets and different algorithms interact to shape model behaviour **[ref: P10]**.
- Working with Mozilla, as part of Mila-Mozilla collaboration, developing efficient memory consolidation algorithms through self-reflection for performant, private, and user-sovereign agent memory systems.

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]**.
- a large-scale image-text dataset across 50 diverse cultures with objectives for inclusive representation learning, and two benchmarks reveals gaps in multicultural understanding in VLMs **[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.
- 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.
- 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
Transactions of the Association for Computational Linguistics 2026
Also presented at CAO Workshop at ICLR'26 **[TACL'26]**

[P9] **Societal Alignment Frameworks Can Improve LLM Alignment** [🔗]
Karolina Stanczak, Nicholas Maede, Mehar Bhatia, Hattie Zhou,..., Sylvie Delacroix, Gillian K. Hadfield, Siva Reddy
ACM Conference on Fairness, Accountability, and Transparency 2026.
Also presented at BiAlign Workshop at ICLR'25 and IASEAI '25 **[FAccT'26]**

[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) **[TMLR'25]**

[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

2025 - 2028	**FRQNT Doctoral Training Research Scholarship** [🔗], awarded funding of $100,000	Quebec, Canada
2025 - 2027	**Mila EDI Scholarship - Women in AI Category** [🔗], awarded funding of $24,000	Mila, Canada
2024	**Prof. Cho Diversity Scholarship**, awarded funding of $1,500	Mila, Canada
2024 - 2028	**Grad Excellence Award in CS**, awarded funding of $6,000 annually	McGill University, Canada
2024	**BPOC Graduate Excellence Award**, awarded funding of $1,650	UBC, Canada
2022 - 2024	**Vector Institute for AI Research Grant**, awarded funding of $4,000 annually	Toronto, Canada
2022 - 2024	**International Tuition Award**, covering full international tuition differential	UBC, Canada
2019	**Honorable Mention at ICDM Conference 2019**, including travel grant	Beijing, China
2016 - 2020	**Institute Merit Scholarship**, awarded to top 10% students in the batch	SNU, India

Invited Talks

April 2026	**Microsoft Research India**: Value Drifts: Tracing Value Alignment During LLM Post-Training [Slides: 🌐]
March 2026	**Guest Lecturer McGill COMP/LING 345**: Social Biases in NLP Systems [Slides: 🌐]
Dec 2025	**Computational Linguistics in Québec Consortium**: Value Drifts: Tracing Value Alignment During LLM Post-Training [Slides: 🌐]
Nov 2025	**UBC Computer Science**: Value Drifts: Tracing Value Alignment During LLM Post-Training [Slides: 🌐]
Aug 2024	**UBC Computer Science**: Exploring Cultural Competence in Language and Multimodal Models [Slides: 🌐]
Feb 2024	**Guest Lecturer UBC CPSC 532V**: Exploring Cross-Cultural Phenomena in NLP [Slides: 🌐]

Teaching Experience

Winter 2026	**Teaching Assistant McGill COMP/LING 345** From Natural Language to Data Science (Gaurav Kamath) [🌐]
Fall 2025	**Teaching Assistant McGill COMP 545** Natural Language Understanding with DL (Prof. Siva Reddy) [🌐]
Winter 2025	**Teaching Assistant McGill COMP/LING 345** From Natural Language to Data Science (Prof. Siva Reddy) [🌐]
April 2023/24	**Teaching Assistant Vector Institute of AI** Prompt Engineering Labs (over 120 participants)
Fall 2022	**Teaching Assistant UBC CPSC 436N** Natural Language Processing (Prof. Vered Shwartz) [🌐]

Media

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

Leadership and Service

Reviewer	EMNLP'26 NeurIPS'26 COLM'26 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