

Jay Gala

Research MSc @ Mila / McGill

[Website](#) [@ Email](#) [GitHub](#) [Google Scholar](#) [Semantic Scholar](#) [LinkedIn](#)

Education

McGill University / Mila

Aug 2025 - Jul 2027 (Expected) | CGPA: **4.00/4.00**

Master of Science (MSc) in Computer Science (Thesis)

Advisors: Siva Reddy, Dzmitry Bahdanau

Representation Learning, Natural Language Processing, Reinforcement Learning, Large Language Models Seminar

Dwarkadas J. Sanghvi College of Engineering (University of Mumbai)

Aug 2017 - Jul 2021 | CGPA: **9.86/10.00**

Bachelor of Engineering (B.E.) in Computer Engineering

Applied Math, Discrete Math, Algorithms, Machine Learning, Artificial Intelligence, Natural Language Processing.

Publications

Complete List at [Google Scholar](#) and [Semantic Scholar](#) (* = equal contribution)

- [12] **BRIDGE: Predicting Human Task Completion Time From Model Performance** [Paper | Code | Website]
Fengyuan Liu*, Jay Gala*, Nilaksh, Dzmitry Bahdanau, Siva Reddy, Hugo Larochelle
International Conference on Machine Learning [ICML 2026]
- [11] **LLMs Can Compensate for Deficiencies in Visual Representations** [Paper]
Sho Takishita*, Jay Gala*, Abdelrahman Mohamed, Kentaro Inui, Yova Kementchedjheva
Findings of the Empirical Methods in Natural Language Processing [Findings - EMNLP 2025]
- [10] **MMTEB: Massive Multilingual Text Embedding Benchmark** [Paper | Code]
Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, et al.
International Conference on Learning Representations [ICLR 2025]
- [9] **SHADES: Towards a Multilingual Assessment of Stereotypes in Large Language Models** [Paper]
Margaret Mitchell, . . . , Jay Gala, . . . , Zeerak Talat
North American Chapter of the Association for Computational Linguistics [NAACL 2025]
- [8] **Leverage Class-Specific Accuracy to Guide Data Generation for Improving Image Classification** [Paper]
Jay Gala, Pengtao Xie
International Conference on Machine Learning [ICML 2024]
- [7] **Critical Learning Periods: Leveraging Early Training Dynamics for Efficient Data Pruning in MT** [Paper]
Everlyn Chimoto, Jay Gala, Orevaoghene Ahia, Julia Kreutzer, Bruce Bassett, Sara Hooker
Findings of the Annual Meeting of the Association for Computational Linguistics [Findings - ACL 2024]
- [6] **An Empirical Study of In-context Learning in LLMs for Machine Translation** [Paper | Code]
Pranjal Chitale*, Jay Gala*, Raj Dabre
Findings of the Annual Meeting of the Association for Computational Linguistics [Findings - ACL 2024]
- [5] **RomanSetu: Efficiently unlocking multilingual capabilities of LLMs via Romanization** [Paper]
Jaavid Aktar Husain, Raj Dabre, Aswanth Kumar, Jay Gala, Thanmay Jayakumar, Ratish Puduppully, Anoop Kunchukuttan
The Annual Meeting of the Association for Computational Linguistics (🏆 **Senior Area Chair Award**) [ACL 2024]
- [4] **CVQA - Culturally-diverse Multilingual Visual Question Answering Benchmark** [Paper | Website]
David Romero, . . . , Jay Gala, . . . , Alham Fikri Aji
Conference on Neural Information Processing Systems Datasets & Benchmark track [NeurIPS 2024]
- [3] **Airavata: Introducing Hindi Instruction-tuned LLM** [Paper | Code]
Jay Gala, Thanmay Jayakumar, . . . , Mitesh Khapra, Raj Dabre, Rudra Murthy, Anoop Kunchukuttan
ArXiv Preprint (Technical Report) [arXiv 2024]
- [2] **IndicTrans2: Towards High-Quality and Accessible MT Models for Indian Languages** [Paper | Code]
Jay Gala*, Pranjal Chitale*, . . . , Mitesh Khapra, Raj Dabre, Anoop Kunchukuttan
Transactions on Machine Learning Research [TMLR 2023]
- [1] **A Federated Approach for Hate Speech Detection** [Paper | Code]
Jay Gala*, Deep Gandhi*, Jash Mehta*, Zeerak Talat
European Chapter of the Association for Computational Linguistics [EAACL 2023]

Experience

Mila - Quebec AI Institute

Aug 2025 - Present

Graduate Student Researcher

Advisors: Siva Reddy, Dzmitry Bahdanau

> Currently working on post-hoc adaptation of pre-trained MoE models to reduce the all-to-all collective communication bottleneck (dispatch/combine) in expert parallelism for efficient multi-node inference deployments.

- › Proposed **BRIDGE** framework to predict human task completion time for AI agents and forecast long-horizon agentic capabilities using Item Response Theory (IRT). We observed a log-linear relationship between IRT-inferred task difficulty and human completion time, requiring no human annotations and only a task-model binary performance matrix for estimation on new benchmarks.
- › Our results on a suite of agentic benchmarks show 92.3% of tasks within a 0.5x–2x tolerance band for human completion time estimates. Forecasting shows task length solvable at 50% accuracy doubles approximately every 6 months, with the best current agent capable of solving tasks up to ~2-2.5 hours in length.

MBZUAI

May 2024 - Jul 2025

Research Associate

Advisors: [Yova Kementchedjhieva](#) and [Alham Fikri Aji](#)

- › Explored retrofitting visual modality knowledge into pre-trained LLMs without an explicit vision encoder. Preliminary experiments showed that the text and image subspaces are largely disjoint with minimal transfer from pre-trained LLMs, particularly for small-scale LLMs (1-3B parameters) under low-data regimes.
- › Investigated the role of language decoders in VLMs like LLaVA ([Liu et al. 2024](#)) using the task of object parts recognition. Through attention-knockout to limit context and logit lens analysis, we show that CLIP image representations encode rich information about object parts effectively extractable via the language decoder.
- › In cases of limited contextualization of object parts from CLIP, LLaVA's language decoder compensates by refining image features, recontextualizing parts, and recovering most of their identifiability.

AI4Bharat (IIT Madras)

Aug 2022 - Apr 2024

AI Resident

Advisors: [Mitesh Khapra](#), [Anoop Kunchukuttan](#) and [Raj Dabre](#)

- › Mined 5M high-quality bitext pairs from the web (ebooks, lecture transcripts, etc) using LaBSE and margin score.
- › Developed SOTA IndicTrans2 translation models and created a challenging IN22 benchmark for 22 Indian languages. Notably, these models are used by the **Supreme Court of India** and **Wikimedia Foundation** ([Coverage](#)).
- › Developed efficient Indic-Indic (non-English) translation models by repurposing components from independently pretrained English-centric translation models. Distilled IndicTrans2 translation models with a **~5x reduction in model size** and **~36% reduced inference time**. Check out the [blog](#) for more details.
- › Study various aspects influencing ICL abilities of LLMs like BLOOM ([Scao et al., 2022](#)) and Llama 2 ([Touvron et al., 2023](#)) for MT task to ascertain if ICL is example-driven or instruction-driven.

Cohere For AI Research Collaboration

Jun 2023 - Feb 2024

Independent Researcher (Remote)

Advisors: [Sara Hooker](#), [Julia Kreutzer](#) and [Bruce Bassett](#)

- › Worked on understanding the effective ways of data pruning for MT by leveraging Checkpoints Across Time (**CAT**).
- › Experimental results demonstrate superior performance using perplexity from early model checkpoints compared to sentence embedding models for high-resource pairs (En-De, En-Fr) and vice-versa for low-resource pairs (En-Sw).

MBZUAI Research Collaboration

Sep 2021 - Dec 2022

Independent Researcher (Remote)

Advisor: [Zeerak Talat](#)

- › Proposed cross-dataset generalization for hate speech detection using Federated Learning extending [Fortuna et al. \(2021\)](#).
- › Experiments show around 10% improvement in F1-score with relatively less data compared to centralized training.

University of California San Diego

Jun 2021 - Jun 2022

Research Intern (Remote)

Advisor: [Pengtao Xie](#)

- › Implementation of [Learning from Mistakes for Neural Architecture Search](#) ([Garg et al., 2021](#)) in PyTorch [[Code](#)].
- › Proposed an efficient multi-level optimization algorithm as an extension to [Garg et al. \(2021\)](#) for improving NAS by conducting performance-aware data generation using class-wise evaluation during the architecture search.
- › Model-agnostic framework that can be coupled with any gradient-based (differentiable) search approaches.

Tata Consultancy Services

Dec 2019 - Feb 2020

Machine Learning Intern

- › Developed models using VAEs and K-means clustering for customer behavior analysis to prevent customer churn.
- › Prepared a custom dataset by developing surveys to handle open-ended and closed-ended questions.
- › Extracted feedback responses from handwritten survey forms using OCR, achieving 12% CER and 18% WER.

Projects

Ocubot - Image-based Dialog [[Report](#) | [Code](#)]

- › Bachelor's project which focused on improving performance on the multimodal task of [Visual Dialog](#).
- › Adversarial analysis of existing systems to identify modality biases towards historical context and salient visual features.
- › Reduced modality biases by improving visual context with dense captions and attention over these captions.
- › Achieved competitive performance to the baseline with around 70% training data (85K images out of 120K images).

Pothole Detection and Depth Estimation [Report | Code]

- › Developed an autonomous surveillance system for road safety to identify potholes using YOLOv4 and estimate the depth and dimensions of the pothole using triangular similarity.
- › Collected and released a dataset of 1.2K pothole images annotated as per the YOLO labeling format.

Anomaly Detection in ECG Signals

- › Industry collaboration to develop neural models for detecting anomalies in processed ECG signals from IoT devices with a human-in-the-loop approach to semi-automate the process while ensuring the safety of human lives.
- › Applied distributed computing algorithms for speed improvements during inference and load balancing by 60%.

Skills

Languages	Python, C, Java, JavaScript, SQL, HTML5, LaTeX
Libraries	PyTorch, Keras, Fairseq, Transformers, Scikit-learn, NumPy, Pandas, OpenCV, SpaCy, NLTK, Flask, FastAPI
Others	Git, Jupyter, Docker, Raspberry Pi

Academic Service

Volunteer	EACL 2023, ACL 2024
Reviewer	EACL 2024, ACL Rolling Review, TMLR, NeurIPS 2025, ICLR 2025, ICLR 2026

Co-Curricular Activities

- › Recipient of the Fonds de recherche du Québec Master's Research (FRQNT) Scholarship for the 2026–2027 year.
- › **Teaching Assistant** for COMP 345 / LING 345: From Natural Language to Data Science, McGill University (Winter 2026).
- › Gave a talk on in-context learning capabilities of LLMs for MT at the SNLP Reading Group, Microsoft Research India.
- › **Presented Tutorial on Developing SOTA MNMT Systems for Related Languages at AACL-IJCNLP 2023.**
- › **Teaching Assistant** for Summer Machine Learning Course, UMLSC 2021, supported by **Google Research India.**
- › Founding member of **SimPPL**, a non-profit research collective mentoring aspiring researchers from Indian educational institutes. Develop Parrot, a tool for auditing online disinformation on social media, in partnership with The Sunday Times and Ippen Digital.