

# Lester James V. Miranda

[ljvmiranda@gmail.com](mailto:ljvmiranda@gmail.com) | [github.com/ljvmiranda921](https://github.com/ljvmiranda921) | [ljvmiranda921.github.io](https://ljvmiranda921.github.io)

## EXPERIENCE

---

**Allen Institute for AI** 2023 – Present  
*Predoctoral Young Investigator* Seattle, WA

- Conducts research broadly in LLM post-training and alignment, such as preference annotation (MULTIPREF), reward model evaluation (REWARDBENCH), and fully open-source post-training recipes (TÜLU 3, OLMo 2).

**Explosion GmbH** 2021 – 2023  
*Machine Learning Engineer, spaCy Team* Berlin, DE

- Authored a technical paper on benchmarking spaCy's hash-embedding trick on span categorization and NER.
- Developed human-in-the-loop and LLM annotation workflows for Prodigy, our data annotation product.
- Built several open-source NLP tools such as `spacy-llm` (production LLM pipelines), `vscode-prodigy` (VSCode extension for data annotation), and `spaCy projects` (end-to-end NLP workflows for production).

**Thinking Machines Data Science** 2018 – 2021  
*Machine Learning Researcher* Metro Manila, PH

- Developed several production-grade NLP applications for a sovereign wealth fund in Singapore.
- Led a team in a large-scale digitization project of government financial statements for The World Bank.
- Led an internal team of six in document processing initiatives to improve our Document AI product.

**Preferred Networks** 2018  
*Research Intern* Tokyo, JP

- Project: Implementing a parallelization feature for ChainerRL to support efficient batch Proximal Policy Optimization (PPO) and Advantage Actor Critic (A2C) for reinforcement learning.

## EDUCATION

---

**Waseda University** 2016 – 2018  
*M.Eng. Information Architecture, Neurocomputing Systems Laboratory* Fukuoka, JP

- Thesis: Autoencoder-based Feature Extraction Techniques for Protein Function Prediction
- Awards: Monbukagakusho (MEXT) Japanese Government Scholarship

**Ateneo de Manila University** 2011 – 2016  
*B.S. Electronics and Communications Engineering, Minor in Philosophy (Cum Laude)* Metro Manila, PH

- Thesis: Appliance Recognition using Hall-Effect Current Sensors for Power Management Systems
- Awards: DOST-SEI Merit Scholarship, Ateneo College Scholarship

## OPEN-SOURCE PROJECTS

---

**calamanCy** | <https://github.com/ljvmiranda921/calamanCy> 2023

- Natural language processing toolkit for building Tagalog pipelines based on spaCy.
- Software paper was published in the NLP-OSS workshop at EMNLP '23.
- Associated NER dataset was published in the SEALP workshop at IJCNLP-AAACL '23.

**PySwarms** | <https://github.com/ljvmiranda921/pyswarms> 2018

- Python-based framework for implementing swarm optimization algorithms.
- Software paper was published in the Journal of Open Source Software (JOSS).
- Has over 1k+ GitHub stars and used by over 300 repositories and packages.

## INVITED TALKS

---

**Artisanal Filipino Resources in the Age of LLMs** | *De la Salle University - Manila* 2024

**Labeling with LLMs** | *University of North Carolina - Charlotte* 2024

**Geospatial Data at Scale with Geomancer** | *Databeers Manila, Google Developer Group Conference* 2019

## SELECTED PUBLICATIONS

---

You can also check my [Google Scholar profile](#) (Scholar ID: 2RtnNKEAAAAJ) for an updated list of my publications. Note: an asterisk (\*) denotes equal or major contributions.

- [1] Nathan Lambert\*, Jacob Morrison\*, Valentina Pyatkin\*, Shengyi Huang\*, Hamish Ivison\*, Faeze Brahman\*, [Lester James V. Miranda\\*](#), Alisa Liu, Nouha Dziri, Xinxu Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tülu 3: Pushing Frontiers in Open Language Model Post-Training. *Preprint*, November 2024.
- [2] [Lester James V. Miranda\\*](#), Yizhong Wang\*, Yanai Elazar, Sachin Kumar, Valentina Pyatkin, Faeze Brahman, Noah A. Smith, Hannaneh Hajishirzi, and Pradeep Dasigi. Hybrid Preferences: Learning to Route Instances for Human vs. AI Feedback. *arXiv*, abs/2410.19133, October 2024.
- [3] Srishti Gureja\*, [Lester James V. Miranda\\*](#), Shayekh Bin Islam\*, Rishabh Maheshwary\*, Drishti Sharma, Gusti Winata, Nathan Lambert, Sebastian Ruder, Sara Hooker, and Marzieh Fadaee. M-REWARDBENCH: Evaluating Reward Models in Multilingual Settings. *arXiv*, abs/2410.15522, October 2024.
- [4] Holy Lovenia\*, Rahmad Mahendra\*, Salsabil Maulana Akbar\*, [Lester James V. Miranda\\*](#), et al. SEACrowd: A Multilingual Multimodal Data Hub and Benchmark Suite for Southeast Asian Languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, November 2024. Empirical Methods in Natural Language Processing.
- [5] Nathan Lambert, Valentina Pyatkin, Jacob Morrison, [Lester James V. Miranda](#), Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. RewardBench: Evaluating Reward Models for Language Modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1755–1797, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.
- [6] [Lester James V. Miranda](#). Allen Institute for AI @ SIGTYP 2024 Shared Task on Word Embedding Evaluation for Ancient and Historical Languages. In *Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 151–159, St Julian’s, Malta, March 2024. Association for Computational Linguistics.
- [7] [Lester James V. Miranda](#). calamanCy: A Tagalog Natural Language Processing Toolkit. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pages 1–7, Singapore, Singapore, December 2023. Empirical Methods in Natural Language Processing.
- [8] [Lester James V. Miranda](#). Developing a Named Entity Recognition Dataset for Tagalog. In *Proceedings of the First Workshop in South East Asian Language Processing*, pages 13–20, Nusa Dua, Bali, Indonesia, November 2023. Association for Computational Linguistics.
- [9] Stephen Mayhew, Terra Blevins, Shuheng Liu, Marek uppa, Hila Gonen, Joseph Marvin Imperial, Brje F. Karlsson, Peiqin Lin, Nikola Ljubei, [LJ Miranda](#), Barbara Plank, Arij Riabi, and Yuval Pinter. Universal NER: A Gold-Standard Multilingual Named Entity Recognition Benchmark. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4322–4337, Mexico City, Mexico, June 2024. Association for Computational Linguistics.
- [10] [Lester James V. Miranda\\*](#), Ákos Kádár\*, Adriane Boyd, Sofie Van Landeghem, Anders Søgaard, and Matthew Honnibal. Multi hash embeddings in spaCy. *arXiv*, abs/2212.09255, December 2022.
- [11] [Lester James V. Miranda](#) and Jinglu Hu. Feature Extraction using a Mutually-Competitive Autoencoder for Protein Function Prediction. In *Proceedings of the IEEE International Conference on System, Man, and Cybernetics (SMC)*. IEEE, October 2018. doi: [10.1109/SMC.2018.00234](#).
- [12] [Lester James V. Miranda](#) and Jinglu Hu. A Deep Learning Approach based on Stacked Denoising Autoencoders for Protein Function Prediction. In *Proceedings of the 42nd IEEE Computer Society Signature Conference on Computers, Software, and Applications (COMPSAC)*. IEEE, July 2018. doi: [10.1109/COMPSAC.2018.00074](#).
- [13] [Lester James V. Miranda](#). PySwarms, a research-toolkit for Particle Swarm Optimization in Python. *Journal of Open Source Software (JOSS)*, 3(433), 2018. doi: [10.21105/joss.00433](#).
- [14] [Lester James V. Miranda\\*](#), Marian Joice Gutierrez\*, Samuel Matthew Dumlao, and Rosula Reyes. Appliance Recognition using Hall-Effect Sensors and k-Nearest Neighbors for Power Management Systems. In *Proceedings of the 2016 IEEE Region 10 Conference 2016 (TENCON)*. IEEE, November 2016. doi: [10.1109/TENCON.2016.7847947](#).