

Machel Reid

Google Scholar: <http://scholar.google.com/citations?user=N8ctPiIAAAAJ>

Website: <https://machelreid.github.io>

Email: machelreid@google.com (Google DeepMind)

RESEARCH / WORK EXPERIENCE

Google DeepMind

Research Scientist

London, UK

Sep 2022 - Present

- Co-led effort for **10M context window** Gemini 1.5 Pro release.^{1,2}
- Led multilingual instruction tuning efforts for PaLM 2/Gemini API.
- Data quality and training efficiency for Gemini post-training (SFT/RLHF and other techniques).
- Built training and evaluation infrastructure used for Gemini post-training efforts. This has now been adopted by teams across Alphabet for training and evaluation of LLMs.

Matsuo Lab, The University of Tokyo

Researcher

Tokyo, JP

November 2020 - Sep 2022

- Working on multilingual pretraining methods, language modeling and text editing
- Teaching Assistant for undergraduate Introduction to Machine Learning (ItML) course.

Research Intern

October 2019 - October 2020

- Worked on representation learning for edits, word embeddings for low resource languages, and definition modeling.

Language Technologies Institute, Carnegie Mellon University

Visiting Student

Pittsburgh, PA

May 2021 - August 2021

- Research internship advised by Professor Graham Neubig.
- Working on methods for modeling of editing processes

PUBLICATIONS

1. **Gemini Team.** Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Technical Report*.
2. **Gemini Team.** Gemini: a family of highly capable multimodal models. *Technical Report*.
3. **Machel Reid**, Mikel Artetxe. On the Role of Parallel Data in Cross-lingual Transfer Learning. *ACL 2023 Findings*.
4. Akari Asai, Sneha Kudugunta, Xinyan Velocity Yu, Terra Blevins, Hila Gonen, **Machel Reid**, Yulia Tsvetkov, Sebastian Ruder, Hannaneh Hajishirzi. BUFFET: Benchmarking Large Language Models for Few-shot Cross-lingual Transfer. *Preprint*.
5. Jonas Pfeiffer, Francesco Piccinno, Massimo Nicosia, Xinyi Wang, **Machel Reid**, Sebastian Ruder. mmT5: Modular Multilingual Pre-Training Solves Source Language Hallucinations. *Preprint*.
6. **Machel Reid**, Vincent J. Hellendoorn, Graham Neubig. DiffusER: Diffusion via Edit-based Reconstruction. *International Conference on Learning Representations 2023. ICLR 2023*.
7. Edison Marrese-Taylor, **Machel Reid**, Alfredo Solano. Edit Aware Representation Learning via Levenshtein Prediction. *Workshop on Insights from Negative Results in NLP, EAACL 2023. May 2022. Association for Computational Linguistics*.
Best Paper Award
8. **Machel Reid**, Victor Zhong, Suchin Gururangan, Luke Zettlemoyer. M2D2: A Massively Multi-domain Domain Dataset. *The 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022). December 2022. Association for Computational Linguistics*.

¹<https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/>

²<https://blog.google/technology/ai/long-context-window-ai-models/>

9. Takeshi Kojima, Shixiang Shane Gu, **Machel Reid**, Yutaka Matsuo, and Yusuke Iwasawa. Large Language Models are Zero Shot Reasoners. *Conference on Neural Information Processing Systems (NeurIPS 2022)*. November 2022.
10. David Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, **Machel Reid**, . . . ,Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya. A Few Thousand Translations Go a Long Way! Leveraging Pre-trained Models for African News Translation. *The 2022 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2022)*. July 2022. Association for Computational Linguistics.
[Oral Presentation](#)
11. Itsuki Okimura, **Machel Reid**, Makoto Kawano and Yutaka Matsuo. On the Impact of Data Augmentation on Downstream Performance in Natural Language Processing. *Workshop on Insights from Negative Results in NLP, ACL 2022*. May 2022. Association for Computational Linguistics.
[Oral Presentation](#) **Best Paper Award**
12. **Machel Reid**, Yutaro Yamada, Shixiang Shane Gu. Can Wikipedia Help Offline Reinforcement Learning? January 2022. *arXiv Preprint*.
13. **Machel Reid** and Graham Neubig. Learning to Model Editing Processes. *Findings of EMNLP 2022*. December 2022. Association for Computational Linguistics.
14. **Machel Reid**, Junjie Hu, Graham Neubig, Yutaka Matsuo. AFROMT: Pretraining Strategies and Reproducible Benchmarks for Translation of 8 African Languages. *The 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021)*. November 2021. Association for Computational Linguistics.
Best Paper Nomination, [Oral Presentation](#)
15. **Machel Reid** and Mikel Artetxe. PARADISE: Exploiting Parallel Data for Multilingual Sequence-to-Sequence Pretraining. *The 2022 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2022)*. July 2022. Association for Computational Linguistics.
16. **Machel Reid** and Victor Zhong. LEWIS: Levenshtein Editing for Unsupervised Text Style Transfer. *Findings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL-IJCNLP 2021 Findings)*. August 2021. Association for Computational Linguistics.
17. **Machel Reid**, Edison Marrese-Taylor and Yutaka Matsuo. Subformer: Exploring Weight Sharing for Parameter Efficiency in Generative Transformers. *Findings of The 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021 Findings)*. November 2021. Association for Computational Linguistics.
18. Francis Zheng, **Machel Reid**, Edison Marrese-Taylor and Yutaka Matsuo. Low-Resource Machine Translation Using Cross-Lingual Language Model Pretraining. *AmericasNLP Workshop, NAACL 2021*. Association for Computational Linguistics.
19. Edison Marrese-Taylor, **Machel Reid** and Yutaka Matsuo. Variational Inference for Learning Representations of Natural Language Edits. *Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)*. February 2021. Presented at the 5th Workshop on Representation Learning for NLP (non-archival), ACL 2020.
20. **Machel Reid**, Edison Marrese-Taylor and Yutaka Matsuo. VCDM: Leveraging Variational Bi-encoding and Deep Contextualized Word Representations for Improved Definition Modeling. *The 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*. November 2020. Association for Computational Linguistics.
21. **Machel Reid**, Edison Marrese-Taylor and Yutaka Matsuo. Combining Pretrained High Resource Embeddings and Subword Representations for Low-Resource Languages. *AfricaNLP Workshop, International Conference of Learning Representations (ICLR20) 2020*

INVITED TALKS

Improving adaptation of multilingual large language models

- Research Talk at *University of Copenhagen* (Host: Constanza Fierro, Desmond Elliott) 05/2023

Editing and Diffusion for Text Generation

- Research Talk at *University of Edinburgh* (Host: Pasquale Minervini) 05/2023
- Research Talk at *Instituto Superior Técnico* (Host: André Martins) 05/2023
- Research Talk at *DeepMind* (Host: Kris Cao) 11/2022

- Research Talk at *NLP Colloquium JP* (Host: Sho Yokoi, Ayana Niwa) 01/2023

Incorporating Text Editing into Natural Language Processing

- Research Talk at the *CMU AI Seminar* (Host: Asher Trockman) 05/2022

Data-centric techniques for improving low-resource multilingual pre-training and benchmarking

- Research Talk at *Google Research* (Host: Clara Rivera) 04/2022
- Research Talk at the *Centro Nacional de Inteligencia Artificial* (Host: Carlos Aspillaga) 12/2022

On leveraging existing resources to develop reproducible benchmarks for African languages

- Research Talk at the *Masakahne Machine Translation Mondays* (Host: Julia Kreutzer) 03/2022

Edit Models & Learning and Understanding Implicit Structure in NLP

- Research Talk at the *The University of Tokyo*. (Host: Yoshihiro Kawahara) 12/2021

PARADISE: Exploiting Parallel Data for Multilingual Sequence-to-Sequence Pretraining

- Research talk at *Facebook NLP Summit* (Host: Abhinav Arora) 10/2021

Learning to Model Editing Processes

- Research talk at the *AIST* (Host: Edison Marrese-Taylor) 08/2021

MENTORING

- **Stephen Kiliu** MS Student, AIMS-AMMI Senegal. (2022-Present) Supervising thesis project for African Masters in Machine Intelligence. Paper accepted to ICLR Tiny Papers track.
- **Itsuki Okimura** BS Student, UTokyo (2021-2022; now MS student) — studying impact of data augmentation techniques on finetuning pre-trained language models (for graduation thesis). Won the best paper at the Insights on Negative Results ACL 2022 workshop for this work.
- **Francis Zheng** MS Student, UTokyo (2021; now PhD student) — Using multilingual pre-training techniques for low-resource machine translation for Americas NLP Shared Task at NAACL 2021.

AWARDS & HONORS

NVIDIA Academic Hardware Grant (PI: Edison Marrese-Taylor) Aug 2021
Awarded single NVIDIA A100 for PCIe for academic research on Edit-Aware Language Models.

Fugaku Project (PI: Aleksandr Drozd) May 2021
Awarded 1 million node hours on Fugaku³ for “Training Novel Types of Large-Scale Language Models: Tuning”.

Masayoshi Son Foundation Scholar July 2020 - Present
Supports those with high aspirations and exceptional talents. Support grants for higher education and research. Grants awarded for Cross-lingual NLP research and for research on hierarchical and multimodal language models (with Yutaro Yamada).

Grand Prize Winner, Rakuten Technology Conference Hackathon October 2018
Developed a prototype matchmaking app between babysitters and working parents; made to combat problems that working women face in Japan today (career or children?). <https://blog.api.rakuten.net/rtc-hackathon-winners/>

VOLUNTEER ACTIVITIES

- **Reviewing** AfricaNLP (2021,2022), ICLR (2022), ICML (2022), NeurIPS (2022), NAACL SRW (2022) ; Secondary: EMNLP (2020,2021), INLG (2020), EACL (2021), NAACL (2021)
- **Volunteer** ICML 2020, ACL 2020, ICLR 2020
- **Rakuten RapidAPI Champions Program**
Worked in the capacity of a technology evangelist (API Champion) at Rakuten RapidAPI. Judged at Junction Tokyo 2019 Hackathon.

SKILLS

Software Python, JavaScript, PyTorch, Fairseq, HF Transformers, Linux, Bash/Zsh, L^AT_EX
Languages English (native), Japanese (advanced; almost native)

³<https://www.fujitsu.com/global/about/innovation/fugaku/>