Large Language Models for Low-Resource Languages: A Plan for Te Reo Māori

Luca Blaauw Fossen*[1]

[1]lf226@students.waikato.ac.nz

Abstract

Large Language Models perform remarkably well on high-resource languages, but lag behind for low-resource and Indigenous languages [1]. This has prompted several language communities to create specialized fine-tuned models for their language. This extended abstract presents an early-stage plan to develop the first sovereign Māori large language model. This plan includes curating high-quality Māori text datasets, constructing culturally relevant benchmarks, and performing continual pre-training and instruction-tuning of open-weight foundation models. This work will be done under Māori expert oversight and community participation from Māori language speakers and iwi (tribes), as well as the CARE principles of Collective Benefit, Authority to Control, Responsibility, and Ethics [2]. At this stage, corpus creation, model choice, and evaluation methods remain under exploration.

1 Motivation & Significance

Large language models have recently become remarkably popular in AI research and society in general, due to their impressive performance across a wide range of NLP tasks. Today, LLMs are usually trained as general-purpose assistants, making them easily accessible to laypeople. However, due to differing amounts of available training data, they exhibit a significant performance difference between high-resource languages (such as English, Chinese, Spanish, etc.) and low-resource languages (including Basque, Te Reo Māori, Sámi, Xhosa, etc.). Preliminary evidence from multilingual translation benchmarks suggests that performance on Māori lags behind English by 10-15 chrF points[3], and some languages lag behind by up to 40 chrF points [4].

This performance gap raises a pertinent concern: As large language models continue to advance, speakers of indigenous and smaller languages risk being left with AI and NLP systems that are much less capable than their high-resource counterparts. Reducing such performance gaps is essential to avoid side-lining indigenous languages in society.

*Supervision from Albert Bifet, Te Taka Keegan, Johan Barthelemy and Samia Touileb.

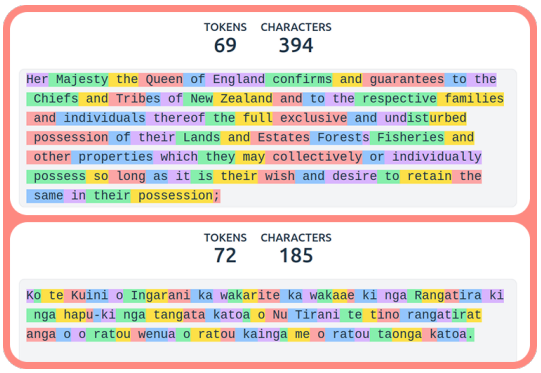

Figure 1. The English and Māori versions of the treaty of Waitangi (Article two, first clause), highlighted by GPT-4 tokenizer. The Māori version shows over 2X higher token density, consuming more tokens despite having less than half the character count of the English version.

2 Related Work

Various local initiatives around the world have emerged to create fine-tuned models tailored to their own languages.

In the Basque country, fine-tuning Llama for the Basque language has produced impressive models competitive with closed-source frontier models across Basque tasks [5], as well as a Basque-adapted evaluation suite called BasqueGLUE [6].

In Singapore, attempts to create models for South East Asian languages have outcompeted GPT-3.5 on South East Asian language tasks [7]. In Norway, NORA.LLM has trained models for Norwegian and Sámi, performing well against similarly sized models [8].

Existing Māori deep learning NLP work focuses on speech, likely due to the prevalence of Māori radio broadcasts and thus audio data. For example, Te Hiku Media have developed ASR technology as well as a speech benchmark for Māori that reflects the language's unique characteristics and diversity of speakers. [9].

To our knowledge, no autoregressive text modeling efforts for Māori have been published, and apart from massively multilingual datasets like FLORES200 [10], there exist no publications on benchmarks designed for Māori LLM evaluation.

Some scholars warn that machine learning may

harm Indigenous languages if applied without community control or linguistic rigor, and can also lead to data poisoning, harming future data collection efforts [11]. We welcome these concerns, and commit to remaining aligned with CARE principles by ensuring Māori authority over data use, community participation in corpus creation and evaluation, and collective benefit through open, culturally governed model development. We also recognize the danger of data poisoning as a real and substantial challenge to low-resource languages, and because of this, we will devote substantial efforts towards developing methods for assessing the quality and authenticity of indigenous text, as well as developing watermarking technologies for our own models.

3 Proposed Work

In order to create an LLM for Māori, we need a high-quality corpus, instruction tuning dataset, and culturally relevant benchmarks. These resources do not yet exist, so an initial step for us is to create them.

3.1 Data Curation

Based on initial corpus analysis, we have identified over 500 million Māori words across public and private datasets. These include news, [12], parliamentary transcripts [13], web corpora like HPLT [14]. However, preliminary filtering suggests that a substantial portion of data may warrant exclusion due to low quality or corpus duplication, so the final usable corpus size is still being determined. Additional private data collections under Māori custodianship such as transcriptions of radio broadcasts, books, and private data collections are also yet to be collected and counted.

We will create instruction fine-tuning datasets directly from Māori speakers through participatory approaches, as well as by adapting existing datasets.

3.2 Data and compute efficiency

Many higher-resource languages have hundreds of billions or even trillions of available tokens. Because Māori resources are comparatively limited, methods of efficient data and compute utilization will be investigated. On the data side, we will explore generating synthetic text from Māori knowledge graphs, as well as common techniques like back-translation and paraphrastic augmentation. On the compute side, we will explore the suitability of second-order optimizers (e.g. Sophia [15]) for language modelling, as well as standard approaches like parameter-efficient fine-tuning (PEFT) and model quantization.

3.3 Tokenizer adaptation

Preliminary experiments indicate that Māori text is tokenized up to twice as inefficiently as English by current models (Figure 1). This presents an opportunity to analyze and adapt tokenizers to reduce token fertility and improve subword coverage for Māori orthography.

3.4 Model evaluation

Model evaluation will rely on both automatic metrics (perplexity, BLEU/chrF++ on custom benchmarks) as well as human assessments of fluency and cultural integrity.

For evaluation, we plan to create MāoriGLUE, a General Language Understanding Evaluation suite created to serve the Māori language community, inspired by the original GLUE [16] and BasqueGLUE [6]. We will consult Te Hiku Media and collaborate with the Māori Language Commission [17] to create this evaluation suite.

We plan to involve the Māori language community through participatory approaches to collect training examples, build benchmarks and perform model evaluation. The entirety of this work will be carried out with guidance from Māori representatives, and will be done in accordance with CARE principles.

4 Expected Contributions

We expect to produce the first sovereign language model for Te Reo Māori. In creating this, we will also produce datasets and other artifacts to further the development of NLP for Māori, as well as results that can inform and guide the development of LLMs and NLP technology for other low-resource languages.

References

[1] K. Dey, P. Tarannum, M. A. Hasan, I. Razzak, and U. Naseem. *Better to Ask in English: Evaluation of Large Language Models on English, Low-resource and Cross-Lingual Settings*. Version Number: 1. 2024. DOI: 10.48550/ARXIV.2410.13153. URL: https://arxiv.org/abs/2410.13153 (visited on 10/17/2025).

[2] S. R. Carroll, I. Garba, O. L. Figueroa-Rodríguez, J. Holbrook, R. Lovett, S. Materechera, M. Parsons, K. Raseroka, D. Rodriguez-Lonebear, R. Rowe, R. Sara, J. D. Walker, J. Anderson, and M. Hudson. "The CARE Principles for Indigenous Data Governance". In: *Data Science Journal* 19 (Nov. 4, 2020), p. 43. ISSN: 1683-1470. DOI: 10.5334/dsj-2020-043. URL: http://datascience.codata.org/articles/10.5334/dsj-2020-043/ (visited on 10/17/2025).

[3] L. Fossen. *Large Language Models for Low-Resource Languages*. AI Institute at Waikato University, Mar. 2024.

[4] N. Robinson, P. Ogayo, D. R. Mortensen, and G. Neubig. "ChatGPT MT: Competitive for High- (but Not Low-) Resource Languages". In: *Proceedings of the Eighth Conference on Machine Translation*. Proceedings of the Eighth Conference on Machine Translation. Singapore: Association for Computational Linguistics, 2023, pp. 392–418. DOI: 10.18653/v1/2023.wmt-1.40. URL: https://aclanthology.org/2023.wmt-1.40 (visited on 10/17/2025).

[5] J. Etxaniz, O. Sainz, N. Miguel, I. Aldabe, G. Rigau, E. Agirre, A. Ormazabal, M. Artetxe, and A. Soroa. "Latxa: An Open Language Model and Evaluation Suite for Basque". In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Bangkok, Thailand: Association for Computational Linguistics, 2024, pp. 14952–14972. DOI: 10.18653/v1/2024.acl-long.799. URL: https://aclanthology.org/2024.acl-long.799 (visited on 10/17/2025).

[6] G. Urbizu, I. San Vicente, X. Saralegi, R. Agerri, and A. Soroa. "BasqueGLUE: A Natural Language Understanding Benchmark for Basque". In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Ed. by N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis. Marseille, France: European Language Resources Association, June 2022, pp. 1603–1612. URL: https://aclanthology.org/2022.lrec-1.172/.

[7] X.-P. Nguyen, W. Zhang, X. Li, M. Aljunied, Z. Hu, C. Shen, Y. K. Chia, X. Li, J. Wang, Q. Tan, L. Cheng, G. Chen, Y. Deng, S. Yang, C. Liu, H. Zhang, and L. Bing. "SeaLLMs - Large Language Models for Southeast Asia". In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations). Bangkok, Thailand: Association for Computational Linguistics, 2024, pp. 294–304. DOI: 10.18653/v1/2024.acl-demos.28. URL: https://aclanthology.org/2024.acl-demos.28 (visited on 10/17/2025).

[8] D. Samuel, V. Mikhailov, E. Velldal, L. Øvrelid, L. G. G. Charpentier, A. Kutuzov, and S. Oepen. "Small Languages, Big Models: A Study of Continual Training on Languages of Norway". In: *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*. Ed. by R. Johansson and S. Stymne. Tallinn, Estonia: University of Tartu Library, Mar. 2025, pp. 573–608. ISBN: 978-9908-53-109-0. URL: https://aclanthology.org/2025.nodalida-1.61/.

[9] G. Leoni, L. Steven, T. Keith, K. Mahelona, P.-L. Jones, and S. Duncan. "Solving Failure Modes in the Creation of Trustworthy Language Technologies". In: *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*. Ed. by M. Melero, S. Sakti, and C. Soria. Torino, Italia: ELRA and ICCL, May 2024, pp. 325–330. URL: https://aclanthology.org/2024.sigul-1.39/.

[10] NLLB Team, M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, A. Sun, S. Wang, G. Wenzek, A. Youngblood, B. Akula, L. Barrault, G. M. Gonzalez, P. Hansanti, J. Hoffman, S. Jarrett, K. R. Sadagopan, D. Rowe, S. Spruit, C. Tran, P. Andrews, N. F. Ayan, S. Bhosale, S. Edunov, A. Fan, C. Gao, V. Goswami, F. Guzmán, P. Koehn, A. Mourachko, C. Ropers, S. Saleem, H. Schwenk, and J. Wang. *No Language Left Behind: Scaling Human-Centered Machine Translation*. Version Number: 3. 2022. DOI: 10.48550/ARXIV.2207.04672. URL: https://arxiv.org/abs/2207.04672 (visited on 10/17/2025).

[11] S. N. Moshagen, L. Antonsen, L. Wiechetek, and T. Trosterud. "Indigenous language technology in the age of machine learning". In: *Acta Borealia* 41.2 (July 2, 2024), pp. 102–116. ISSN: 0800-3831, 1503-111X. DOI: 10.1080/08003831.2024.2410124. URL: https://www.tandfonline.com/doi/full/10.1080/08003831.2024.2410124 (visited on 10/17/2025).

[12] *Niupepa: Māori Newspapers*. URL: https://www.greenstone.org/greenstone3/library/collection/niupepa/page/about.

[13] New Zealand Parliament. *Hansard Debates*. 2025. URL: https://www.parliament.nz/en/pb/hansard-debates/.

[14] HPLT Project. *High Performance Language Technologies*. URL: https://hplt-project.org/.

[15] H. Liu, Z. Li, D. Hall, P. Liang, and T. Ma. *Sophia: A Scalable Stochastic Second-order Optimizer for Language Model Pre-training*. Version Number: 4. 2023. DOI: 10.48550/ARXIV.2305.14342. URL: https://arxiv.org/abs/2305.14342 (visited on 10/17/2025).

[16] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman. "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding". In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 353–355. DOI: 10.18653/v1/W18-5446. URL: http://aclweb.org/anthology/W18-5446 (visited on 10/17/2025).

[17] L. C. Maori. *Māori Language Commision*. Te Taura Whiri I Te Reo Māori. URL: https://www.tetaurawhiri.govt.nz/.