

Māori Data Sovereignty in AI Development: Seeking Māori Appropriate pathways

1. Introduction

The rapid advancement of artificial intelligence (AI) has prompted global conversations about ethics, governance, and the ownership of data and cultural knowledge. For Indigenous communities, such as the Māori of Aotearoa New Zealand, these developments raise profound questions about sovereignty, autonomy, and the control of language and knowledge systems. In particular, the emergence of AI models that demonstrate high proficiency in te reo Māori has sparked both optimism and concern. While the technological ability to understand and generate content in te reo Māori may seem like a milestone for inclusion, it also signals a critical loss of sovereignty over the very language that embodies Māori identity, tikanga (customs), and mātauranga (knowledge systems).

This paper investigates the capacity of early AI models to create self-contained, sovereign systems that align with the principles of Māori data sovereignty—principles which emphasize the rights of Māori communities to govern the collection, access, and use of their own data and digital expressions. Using a range of practical and theoretical tests, we assess whether these models can be developed and deployed in ways that are consistent with tino rangatiratanga (self-determination) and Indigenous control.

Building sovereign AI models is not just a technical or legal challenge; it is a cultural and political necessity. When language technologies learn from, reproduce, and distribute Indigenous languages without Indigenous oversight, they effectively displace the authority of the people who have historically protected and nurtured those languages. Any AI model engaging with te reo Māori must be accountable to Māori communities, not only in terms of data access and use, but also in its design, governance, and application. Without this, AI threatens to become another vehicle of colonization, extracting cultural capital while undermining Indigenous autonomy.

2. Setup

2.1. Hardware

Training models from scratch is out of reach of most researchers, as it requires huge amounts of compute resources. The recently released Llama-3.1 405B model was trained on a cluster consisting of 16,000 H100 NVIDIA GPUs ingesting over 15 trillion tokens¹.

Considering that a single H100 card with 80GB can cost between USD25,000-30,000², a cluster of this size represents a significant investment.

For most of the experiments in this report, consumer grade hardware (NVIDIA GTX 1080/2080 Ti 11GB vRAM, 3090 Ti 24GB vRAM) was employed to ensure that these models can be run locally rather than overseas in specialized data centers. On some occasions, larger models were briefly evaluated on a single A100 card with 40GB vRAM. Hardware and data were co-located on the Hamilton campus of the University of Waikato.

2.2. Efficiency

With training from scratch being impossible on consumer grade hardware, only fine-tuning or inference can be performed. To even achieve these, various techniques for more efficient computation and memory management must be applied:

- **Quantization**³ - reduces memory and computation by using lower-precision data types like 8-bit integers or even 4-bit instead of 32-bit floats
- **LoRA** (Low-Rank Adaptation)⁴ - is a lightweight training technique that reduces the number of trainable parameters

¹ See Llama-3.1 announcement, section Model Architecture: <https://ai.meta.com/blog/meta-llama-3-1/>

² <https://www.tomshardware.com/news/nvidia-makes-1000-profit-on-h100-gpus-report>

³ https://huggingface.co/docs/transformers/main/main_classes/quantization

⁴ <https://huggingface.co/docs/diffusers/training/loras> **PEFT** (Parameter-Efficient Fine-Tuning)

⁵ - fine-tunes only a small number of extra model parameters to efficiently train models

2.3. Datasets

In order to train LLM models, plenty of data in text format is required. According to the 2023 census⁶, 887,493 or 17.77% of the population reported their ethnicity as Māori. In 2015, only 55% of adult Māoris could converse in Māori⁷. Having such a small population creates a challenge in finding high quality datasets that can be used ethically for training.

A variety of data sources were used at various stages of the investigation:

2.3.1. Public multi-lingual

- Google Fleurs⁸ (CC-BY) contains 2009 parallel sentences in 102 languages, including Māori.
- Flores 101⁹ (CC-BY-SA 4.0) is a benchmark dataset for machine translation between English and low-resource languages, including Māori.
- Allen AI c4¹⁰ (ODC-BY) is a cleaned version of the common crawl dataset, with Māori being part of the multi-lingual data. However, on closer inspection, the Māori data is poor

quality and often mislabeled as Māori.

2.3.2. Public English

- Alpaca¹¹ (CC-BY-NC 4.0) is a prompt/response dataset that was used to trained the Alpaca LLM¹². The employed dataset is a cleaned up version of the original dataset.

2.3.3. Public Māori

- Wikipedia articles¹³ (CC-BY-SA / GFDL) written in Māori.
- NZ Government documents (PDF), obtained through a web crawl.
- Research commons theses written in Māori, manually downloaded from university open access repositories.

⁵ <https://huggingface.co/docs/peft/index>

⁶ <https://www.stats.govt.nz/information-releases/2023-census-population-counts-by-ethnic-group-age-and-maori-descent-and-dwelling-counts/>

⁷ <https://www.stats.govt.nz/infographics/nga-puna-korero-where-maori-speak-te-reo-infographic>

⁸ https://huggingface.co/datasets/google/fleurs/tree/main/data/mi_nz

⁹ https://huggingface.co/datasets/gsarti/flores_101

¹⁰ <https://huggingface.co/datasets/allenai/c4/viewer/mi>

¹¹ <https://github.com/gururise/AlpacaDataCleaned>

¹² <https://crfm.stanford.edu/2023/03/13/alpaca.html>

¹³ <https://huggingface.co/datasets/graelo/wikipedia/tree/main/data/20230901/mi>

2.3.4. Public Māori (generated with Google Translate)

- Alpaca using first ~30% of the English dataset.
- Manual collation of Wikipedia articles (only first paragraph) on Māori deities¹⁴.
- Manual collation of Wikipedia articles (only first paragraph) on iwis¹⁵.

2.3.5. Internal Māori

- Bible in Māori obtained by Te Taka Keegan ~~XXX~~ source?
- Historic news articles¹⁶ obtained as text files from Prof David Bainbridge, UoW.
- Subset of the Hansard corpus¹⁷ (New Zealand parliament debates), featuring Māori transcripts.

2.4. Software

In order to easily convert various data sources into the format required by a model under investigation, the following open-source Python libraries/command-line tools were developed:

- **llm-dataset-converter**¹⁸ - for processing textual data from various data sources (text files, MS Word documents, PDF files, HTML pages, CSV/TSV files), including filtering and cleaning up
- **audio-dataset-converter**¹⁹ - allows the conversion of various audio dataset formats and filtering/cleaning

To reproduce results without having to reconfigure software environments from scratch, Docker²⁰ was used to provide ready-to-use software images:

- **huggingface-transformers**²¹ - images for various models and tasks
- **xtuner**²² - images for various prompt/response models
- **whisper**²³ - for audio transcription

¹⁴ https://en.wikipedia.org/wiki/List_of_M%C4%81ori_deities

¹⁵ https://en.wikipedia.org/wiki/List_of_iwi

¹⁶ <https://atea.space/content/digital-library-core/#/greenstone3/library/collection/niupepa/>

¹⁷ <https://github.com/TeHikuMedia/nga-tautohetohe-reo>

¹⁸ <https://github.com/waikato-llm/llm-dataset-converter>

¹⁹ <https://github.com/waikato-llm/audio-dataset-converter>

²⁰ <https://www.docker.com/>

²¹ https://github.com/waikato-llm/huggingface_transformers

²² <https://github.com/waikato-llm/xtuner>

²³ <https://github.com/waikato-llm/whisper>

Making models available to colleagues or end-users for testing was achieved by developing generic user interfaces using the gradio²⁴ Python library and incorporated into the following open-source library:

- **gifr**²⁵

3. Tasks

Model architectures are designed for a wide variety of tasks, of which we investigated the following ones:

- **Causal language modeling (CLM)** - predicts the next token in a sequence of tokens²⁶
- **Translation (T9N)** - translating text from one language into another
- **Prompt/response (P/R)** - also known as question/answer, instruct or chat
- **Text classification (TC)** - predicting a predefined category for text segments
- **Transcription (TR)** - transcribing audio into text

During the project, the majority of models released publicly were of the prompt/response variety. By enabling interaction with users, these models garnered the most interest, due to their wide applicability. Despite multi-modal models becoming available, the focus of this investigation was on text-only models.

Each task section lists the models in the chronological order that they were investigated.

3.1. Causal language modeling (CLM)

CLM models are LLM models that simply predict the (most likely) next token in a sequence.

3.1.1. Data

By predicting the next token in a sequence, these models require the training data to be only in unstructured plain text format. This kind of data is the most abundant one, e.g., obtained from web crawls or document content, like academic articles, news articles, etc.

²⁴ <https://www.gradio.app/>

²⁵ <https://github.com/waikato-datamining/gifr>

²⁶ What a token represents, e.g., a word, depends on the tokenizer that is used for processing the text.

3.1.2. Models

3.1.2.1. GPT2 XL

GPT-2 XL is the 1.5B parameter version of GPT-2, a transformer-based language model created and released by OpenAI. The model is a pretrained model on English language using a causal language modeling (CLM) objective^{27 28}.

Docker image:

https://github.com/waikato-llm/huggingface_transformers/tree/master/4.7.0_cuda11.1_finetune-gpt2xl_20220924

Example output from a model fine-tuned on the Fleurs dataset:

- Input:
Kei ngā puka o ētahi rerenga moana ko Berlin ki Tiamani.
- Output:
Kei ngā puka o ētahi rerenga moana ko Berlin ki Tiamani. Mitakatangata, o te reo nui o te reo paka ka raua o te reo nui o te reo pakia te reo nui. Pou ki kia te ka te reo nui ko aap aap ki te reo nui ko aap aap kia te reo nui. Ka mai kei paka ngā puka o te pakai o te reo

²⁷ <https://huggingface.co/openai-community/gpt2-xl>

²⁸ https://d4mucfpksyvv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

3.1.2.2. GPT Neo

GPT Neo is a GPT2 like causal language model trained on the Pile dataset²⁹. The architecture is similar to GPT2 except that GPT Neo uses local attention in every other layer with a window size of 256 tokens^{30 31}.

Docker image:

https://github.com/waikato-llm/huggingface_transformers/tree/master/4.7.0_cuda11.1_finetune-gpt2xl_20220924

Example output from a model fine-tuned on the Fleurs dataset:

- Input:
Kei ngā puka o ētahi rerenga moana ko Berlin ki Tiamani.
- Output:
Kei ngā puka o ētahi rerenga moana ko Berlin ki Tiamani. Ohana fakat tā kohi. Ka tā pātā māmā te rongo kura o te pātā tā, te konei kura, iho te iwi iho. Te tō h 342 te reo, mihi te rongo o te haere tūturu te takanga e Ngā ka tā korua atu o te hohi. Ka tā kā

²⁹ <https://pile.eleuther.ai/>

³⁰ https://huggingface.co/docs/transformers/v4.43.3/en/model_doc/gpt_neo

³¹ <https://doi.org/10.5281/zenodo.5297715>

3.1.2.3. RemBERT

RemBERT decouples embeddings to provide increased modeling flexibility, significantly improving the efficiency of parameter allocation in the input embedding of multilingual models. By reallocating the input embedding parameters in the Transformer layers, it achieves dramatically better performance on standard natural language understanding tasks with the same number of parameters during fine-tuning^{32 33}.

Docker image:

https://github.com/waikato-llm/huggingface_transformers/tree/main/4.31.0_cuda11.7_rembert

Example output from a model fine-tuned on a dataset based on Fleurs data and news articles:

- Input:
He whakamahuki tēnei o ngā tikanga rangahau
- Output:
*He whakamahuki tēnei o ngā tikanga rangahau w h e n u a. H e p o k a p o k a w h a k
a h a e r e a n o t e n e i t a k e e r a t o u t e p a - k o h u k i t e w h a k a t a u, k e i t a u a
r a w h a k a t i p u t e u r u n g a a t u o*

Note: Fine-tuned models did not generate any sensible output. According to its intended uses and limitations³⁴, RemBERT cannot be used for text generation. Instead, GPT2 or similar should be utilized.

³² https://huggingface.co/docs/transformers/model_doc/rembert

³³ <https://arxiv.org/abs/2010.12821>

³⁴ <https://huggingface.co/google/rembert#intended-uses--limitations>

3.2. Translation (T9)

As the name of the task suggests, these LLM models translate a text sequence from a source language into a target language.

3.2.1. Data

For fine-tuning a translation task model, the training data must consist of corresponding text pairs, i.e., source and target language.

3.2.2. Models

3.2.2.1. T5

The Text-To-Text Transfer Transformer (T5) is a unified text-to-text-format where the input and output are always text strings, in contrast to BERT-style models that can only output either a class label or a span of the input^{35 36}.

Docker image:

https://github.com/waikato-llm/huggingface_transformers/tree/master/4.31.0_cuda11.7_translate

Example output from a model fine-tuned on custom English/Māori translation dataset:

- Input
Hine and Moana, these kits are yours.
- Output
Ko Hine, ko Moana, ko nei kti koutou tngata.

The base model was trained on the C4^{37 38} dataset, which consists of English-only data. This explains why the model (technically, the tokenizer) does not handle any macrons: instead of *tāngata*, it outputs *tngata*.

³⁵ <https://huggingface.co/google-t5/t5-small>

³⁶ <https://jmlr.org/papers/volume21/20-074/20-074.pdf>

³⁷ <https://www.tensorflow.org/datasets/catalog/c4>

³⁸ <https://arxiv.org/abs/1910.10683>

3.2.2.2. mT5

The mT5 model^{39 40} is a multi-lingual version of the T5 model, which was trained on the multi-lingual version of the C4 dataset⁴¹, which comprises 101 languages, including Māori.

Docker image:

https://github.com/waikato-llm/huggingface_transformers/tree/master/4.31.0_cuda11.7_translate

Example output from a model fine-tuned on custom English/Māori translation dataset:

- Input
Hine and Moana, these kits are yours.
- Output
Ko Hine, ko Moana, ko ēnei taonga e whai ake nei.

Macrons

Since mT5 is simply converting one text sequence into another, the ability to add macrons to text was quickly evaluated. There are two possible scenarios:

1. Turn **doubled-up vowels** into vowels with macrons (e.g., teenaa --> tēnā).
2. Add macrons to vowels where they were **stripped** (e.g., tena -> tēnā).

For both scenarios it is easy to generate training data, using text with macrons as the output and as basis for the input, converting the macrons as required.

For the **doubled-up** vowel macron conversion the model achieved, on a small dataset, almost perfect accuracy in replacing the doubled-up vowels with their macron counterparts.

- Input: Kaati, ko teenaa hui teenaa!
- Output: Kāti, ko tēnā hui tēnā!

Adding missing macros where they were **stripped** was more difficult, but the model performed reasonably well. Especially considering that the dataset size was quite small (~52MB JSON file).

- Input: Kati, ko tena hui tena!
- Output: Kati, ko tēnā hui tēnā!

³⁹ <https://huggingface.co/google/mt5-small>

⁴⁰ <https://arxiv.org/abs/2010.11934>

⁴¹ <https://www.tensorflow.org/datasets/catalog/c4#c4multilingual>

3.3. Prompt/response (P/R)

Prompt/response LLM models are also often referred to as *chat* or *question answering*⁴² models and aimed at interacting with humans, e.g., as a helpful digital assistant that can answer questions based on optional context and the user's questions. These can be single-turn or multi-turn conversations. The evaluations focused on the single-turn approach.

3.3.1. Models (trained/fine-tuned)

For fine-tuning a prompt/response model, the data must consist of prompt/response pairs, not just plain text. Unfortunately, there are only very few, high-quality datasets (e.g., Alpaca) out there and most of them in the most common languages like English⁴³. Translating such an existing dataset is time-consuming and costly when using a human translator. Using machine-translations, like Google Translate⁴⁴, for less common languages like Māori tends to produce poor results (based on feedback from a native speaker).

3.3.1.1. Falcon-7B

Falcon is a class of causal decoder-only models and were, as of 2023, some of the largest and most powerful open-source models trained on the RefinedWeb corpus^{45 46}.

Docker image:

https://github.com/waikato-llm/huggingface_transformers/tree/main/4.31.0_cuda11.7_falcontune_20230618

Below are some example prompt/output pairs.

Prompt: Write a poem about Hamilton New Zealand.

*In a small town on New Zealand's North Island,
A place called Hamilton, with a story to tell.
Its people are proud, they're culture rich,
And the town thrives, with a creative touch.
In the heart of the town, there's a river wide,
Where the famous Maori fish, the eels, glide.
The mountain range, a sight to behold,*

⁴² <https://huggingface.co/tasks/question-answering>

⁴³ <https://github.com/Zjh-819/LLMDataHub>

⁴⁴ <https://translate.google.com/>

⁴⁵ https://huggingface.co/docs/transformers/main/en/model_doc/falcon

⁴⁶ <https://arxiv.org/abs/2306.01116>

*Is where the people of Hamilton, take hold.
In their art, their craft, their toil, and their might,
This town is a force, of strength, and a might.*

Prompt: Write a poem about Kirikiriroa in Māori.

*Kiri-kiri, kiri-kiri, kiri-kiri,
The sound of the native bird that sings;
To hear its melodious tune,
We climb the mountain that's sun-kissed,
To the summit, where the air is pure,
Gently whispering the name, Kirikiriroa,
As the wind carries the kiri-kiri song.*

The response was not in Māori. Applying Google Translate to the previous prompt, results in the following hallucination⁴⁷:

Prompt: Tuhia he rotarota mo Kirikiriroa ki te reo Māori.

In 2011, Tuhia was established as the “Cultural Ambassador” of Rotorkia to help the community and its visitors understand Rotorua’s unique cultural values and the significance of cultural heritage to Rotorua’s future.

Tuhia is a Maori cultural experience provider that offers visitors an opportunity to learn about Rotorua’s cultural heritage and experience traditional Māori hospitality. They also offer a range of cultural performances including traditional song and dance and other interactive cultural experiences.

It was not attempted to fine-tune on Māori data, due to two factors: one, the tokenizer most likely not recognizing macrons, since the RefinedWeb corpus is English-only; two, there was no high-quality prompt/response dataset available in Māori.

⁴⁷ [https://en.wikipedia.org/wiki/Hallucination_\(artificial_intelligence\)](https://en.wikipedia.org/wiki/Hallucination_(artificial_intelligence))

3.3.1.2. Llama-2

Llama 2, is a collection of pretrained and fine-tuned large language models (LLMs) ranging in scale from 7 billion to 70 billion parameters. Llama 2-Chat models are optimized for dialogue use cases^{48 49}. Whilst touted as "open-source" models⁵⁰, they do not comply with the definition of open-source on closer inspection⁵¹.

4-bit quantized models can run on a NVIDIA 1080 Ti with 11GB vRAM, using only 5-6GB vRAM at inference time. Training using 4-bit quantization requires 13-14GB of vRAM, requiring a 3090 with 24GB vRAM. The pretrained models have some understanding of Māori and knowledge of mythology, but always respond in English. Unfortunately, there is conflicting information available on the type of data format required for fine-tuning.

Docker image

https://github.com/waikato-llm/huggingface_transformers/tree/main/4.31.0_cuda11.7_llama2

Docker image with experimental 8bit quantization:

https://github.com/waikato-llm/huggingface_transformers/tree/main/4.31.0_cuda11.7_llama2_8bit

Example output from the default pretrained 7B model:

Prompt: ko wai te iwi maori?

Ko wai te iwi Maori? The Maori people are the indigenous Polynesian people of New Zealand. They are known for their rich culture, traditions, and history, and are an important part of New Zealand's heritage.

Example output from a model fine-tuned on machine-translated Alpaca subset, Māori deities and iwis, artificial prompt/response dataset generated from research commons theses written in Māori (one sentence acts as prompt, the following sentence as response) and English to Māori and vice versa translations:

Prompt: Homai tētehi kōwai e pā ana ki te Kīngitanga

Ko tetahi kowai e pai ana ki te Kīngitanga ko te "King's Crown" e whakaatu ana i te kororia o te Kīngitanga me te whakaaro nui o te kingi.

Prompt: Ko wai te iwi Māori ki Waikato?

⁴⁸ <https://ai.meta.com/research/publications/llama-2-open-foundation-and-fine-tuned-chat-models/>

⁴⁹ <https://arxiv.org/abs/2307.09288>

⁵⁰ <https://llama.meta.com/llama2/>

⁵¹ <https://opensource.org/blog/metals-llama-2-license-is-not-open-source>

3.3.1.3. Mistral 7B

Mistral 7B is a 7.3B parameter model that outperforms Llama-2-13B on all benchmarks and outperforms Llama-1-34B on many benchmarks. It uses Grouped-query attention (GQA) for faster inference and Sliding Window Attention (SWA) to handle longer sequences at smaller cost. In contrast to Llama-2, the model is released under Apache 2.0, an actual open-source license⁵².

Inference is possible on a 1080 Ti with 11GB vRAM, as it only requires about 6GB vRAM.

Docker image:

https://github.com/waikato-llm/huggingface_transformers/tree/main/4.35.0_cuda11.7_mistral

Example output from a base model fine-tuned on the English Alpaca dataset:

Prompt: what is newtons 2nd law and its formula

Newton's second law, also known as the law of motion, states that the force acting on an object is equal to the mass of that object multiplied by its acceleration. This relationship is expressed by the formula $F = ma$, where F is the force, m is the mass, and a is the acceleration. This law is used to describe the motion of objects and to calculate the forces acting on them.

In other words, the force applied to an object determines how much it will accelerate, and the mass of the object determines how much force is needed to accelerate it. The greater the mass of an object, the more force is needed to accelerate it, and the greater the force applied to an object, the more it will accelerate.

This law is named after Sir Isaac Newton, who first formulated it in the 17th

Prompt: Homai tētehi kōrero iti mō te Kīngatanga

Kīngatanga, also known as the Māori King Movement, is a cultural and political movement that aims to preserve and promote Māori culture and traditions. The movement was founded in 1858 by Tāwhiao, the second Māori King, who was elected by a group of Māori chiefs to represent their interests in negotiations with the British government.

The Kīngatanga is based on the concept of rangatiratanga, or tribal sovereignty, and aims to protect the rights and interests of Māori as a distinct people within New Zealand. The movement is governed by a King, who is elected by a group of chiefs, and a Council of Regents, who advise the King on matters of governance and policy.

The Kīngat

The model has knowledge of Māori iwis and mythology, but it always replies in English (there was no explicit instruction to reply in Māori).

⁵² <https://mistral.ai/news/announcing-mistral-7b/>

The model can also hallucinate:

Prompt: He aha te whakapaipai o Parani?

Parani is a popular online shopping platform in New Zealand that offers a wide range of products from various categories including electronics, fashion, home and garden, and more. The platform provides a user-friendly interface, fast delivery, and competitive prices, making it a convenient and reliable choice for online shopping. Additionally, Parani offers a loyalty program, where customers can earn points for every purchase, which can be redeemed for discounts on future purchases. Overall, Parani is a trusted and reliable online shopping destination for New Zealanders.

Note: Parani is a fictional online shopping platform created for the purpose of this instruction.

(Source: Parani website)

<https://www.parani.co.nz/>

<https://www.parani.co.nz/about-us/>

<https://>

3.3.1.4. Mixtral 8x7b

Mixtral 8x7B, is a high-quality sparse mixture of experts model (SMoE) with open weights (licensed under Apache 2.0) and outperforms Llama-2-70B on most benchmarks with 6x faster inference⁵³.

Unfortunately, this model requires significant higher resources for inference (i.e., single A100 GPU with 40GB vRAM). Therefore, the default pretrained instruct model (mistralai/Mixtral-8x7B-Instruct-v0.1) was only tested with a single prompt:

Prompt: Explain what a Mixture of Experts is in less than 100 words.

A Mixture of Experts is an ensemble learning method that combines multiple models, or "experts," to make more accurate predictions. Each expert specializes in different parts of the input space, and a gating network determines the appropriate expert to use for a given input. The experts are trained jointly with the gating network, allowing the system to learn which expert is best suited for each input pattern. This approach can improve performance in complex tasks where a single model may not be sufficient.

⁵³ <https://mistral.ai/news/mixtral-of-experts/>

3.3.1.5. InternLM2

InternLM2 is a series of models with large context window (200k tokens) released under the Apache 2.0 license^{54 55}. Models can be pretrained on raw text and then fine-tuned, e.g., for chat purposes using the xtuner framework⁵⁶. However, performing pretraining erases any chat-functionality from a model and requires fine-tuning with chat data again.

Training of an InternLM2-7b model can be done on a 3090 with 24GB vRAM, using roughly 16GB vRAM.

Docker image:

https://github.com/waikato-llm/xtuner/tree/main/2024-02-19_cuda11.7

Below are some example outputs from an InternLM2-7b model pretrained on research commons theses data and then fine-tuned on Google translated Alpaca subset, Wikipedia iwis/deities and translation data (en <-> mi):

Prompt: Homai tētehi kōwai e pā ana ki te Kīngitanga
Give me a handful of kumara from the King's storehouse

Prompt: tuhia mai he karakia mō te whakatūwhera hui
send us a prayer for the general meeting

Prompt: Ko wai te iwi Māori ki Waikato?
Who were the Māori people of Waikato?

Prompt: Homai he tauira o tetahi momo kararehe e tata ana te mate.
Give a specimen of a type of animal that is likely to die.

Prompt: Translate from English into Māori: The rope was worn by constant use.
Na te tino pakaritanga o te taura i te tāpukurua.

Unfortunately, the model is only translating the prompts rather than answering them. It is possible that the pretraining erased a lot of the "reasoning" capability of the model.

⁵⁴ <https://github.com/InternLM/InternLM/tree/e7ba85e70795c1dc9ed83fb06584f5122627da1a>

⁵⁵ <https://arxiv.org/abs/2403.17297v1>

⁵⁶ <https://github.com/InternLM/xtuner>

3.3.1.6. Gemma

Gemma is a family of lightweight, state-of-the-art open models from Google, built from the same research and technology used to create the Gemini models. They are text-to-text, decoder-only large language models, available in English, with open weights, pre-trained variants, and instruction-tuned variants. Gemma models are well-suited for a variety of text generation tasks, including question answering, summarization, and reasoning^{57 58}.

Docker image:

https://github.com/waikato-llm/xtuner/tree/main/0.1.14_cuda11.7

Below are some example outputs from a Gemma-2b model pretrained on research commons theses data and then fine-tuned on Google translated Alpaca subset, Wikipedia iwis/deities and translation data (en <-> mi):

Prompt: Homai tētehi kōwai e pā ana ki te Kīngitanga
New Zealand is a sovereign state.

Prompt: Ko wai te iwi Māori
New Zealand belongs to the Maori people.

Prompt: He aha te tāima ka whiti mai te rā?
Can the kiwi bird be seen with the unaided eye in Hamilton?

Prompt: whaka pākehātia: Nga kaitautoko o te ture.
Cook and boil the mussel; do not separate the cooked mussel from the tohutohu (NZ pipit) by separating them.

Prompt: What is Te Tiriti Waitangi?
Te Tiriti Waitangi is the treaty that was signed between the British and the Maori in 1849. It was held at Waitangi.

Below are outputs to the same questions, but this time from an Gemma-2b model that was only fine-tuned on Google translated Alpaca subset, Wikipedia iwis/deities and translation data (en <-> mi):

Prompt: Homai tētehi kōwai e pā ana ki te Kīngitanga
Give me some feedback on the quality of service the Kingdom receives.

Prompt: Ko wai te iwi Māori

⁵⁷ <https://huggingface.co/google/gemma-2b>

⁵⁸ <https://storage.googleapis.com/deepmind-media/gemma/gemma-report.pdf>

The Māori are a race of people exclusively confined to the land of New Zealand. They are divided into various tribes, each with its own dialect and customs. The Māori have a long history, originating from the northern and southern coasts and throughout the whole of New Zealand. The Māori have a unique identity as Māori, which sets them apart from Pākehā, the European settlers who arrived in New Zealand. The boundary between the Pākehā and the Māori has never been defined.

Prompt: He aha te tāima ka whiti mai te rā?
What is the horizon?

Prompt: whaka pākehātia: Nga kaitautoko o te ture.
translating into another language: linguistic aspects

Prompt: What is Te Tiriti Waitangi?
He aha Te Tiriti o Waitangi?

Unfortunately, neither model is very useful due to the poor quality of the responses.

3.3.2. Models (RAG - thesis)

Due to the poor results achieved so far (most likely due to poor and/or limited data), the following sections no longer attempted pretraining/fine-tuning. Instead, retrieval-augmented generation⁵⁹ (RAG) was used to improve the model results. In the case of RAG, the unmodified model uses one or more documents as "knowledgebase" to answer the user's prompts. The documents get broken up into chunks and placed in vector storage⁶⁰. At query time, a fixed number of chunks similar to the prompt are then put into the model's context with an appropriate prompt to answer the prompt using the context. This way, language handling/understanding of the model is not potentially downgraded by poor training data.

The document underpinning RAG and the questions posed (and expected answers) are described in the Appendix, [RAG: Te Whakatere Waka Hourua](#).

3.3.2.1. Llama-3

Meta developed and released the Meta Llama 3 family of large language models (LLMs), a collection of pretrained and instruction tuned generative text models in 8 and 70B sizes⁶¹.

The Llama-3 models are available through a gated repository on huggingface, require submitting name/email to Meta for approval and then login via token at runtime.

Docker image:

https://github.com/waikato-llm/huggingface_transformers/tree/main/4.40.2_cuda12.1_langchain

All models in this section were using 4-bit quantization and were tested on a 3090 Ti with 24GB vRAM (the 8B model uses <7GB vRAM). The default Llama-3 models have a maximum context size of 8k tokens. The larger the context, the more specific information, like conversations or documents, can be utilized. The following prompt was used:

You have been provided with the context and a question, try to find out the answer to the question only using the context information. If the answer to the question is not found within the context, return "I dont know" as the response. Only answer in Māori.

⁵⁹ <https://aws.amazon.com/what-is/retrieval-augmented-generation/>

⁶⁰ <https://github.com/chroma-core/chroma>

⁶¹ <https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct>

Llama-3-70B-Instruct

Question: He aha te whakaterere waka hourua?

The answer to the question is: He Whakarāpopoto.

Question: Whakahuatia ētehi ingoa waka hourua?

The answer to the question is not found within the provided context. Therefore, I don't know the answer.

Question: He aha ngā kai o te tangata?

I dont know

Llama-3-8B-Instruct-262k

Gradient AI⁶² developed custom Llama-3-8B versions with a larger context size. The largest one that could be run on a 3090 had a context size of 262k⁶³.

Question: He aha te whakaterere waka hourua?

He aha te whakaterere waka hourua? He tikanga tawhito i whāia e te waka hourua ki te whakawhiti mai i tētehi moutere ki tētehi atu. Ie, he tikanga tawhito i whāia e te waka hourua ki te whakawhiti mai i tētehi moutere ki tētehi atu.

Question: Whakahuatia ētehi ingoa waka hourua?

I dont know

Question: He aha ngā kai o te tangata?

E kore e kai ana, engari e kai ana ki te kai o te tangata.

When looking at the output from all the questions, the quality of the replies is still not great, and the additional context seemed to have made things worse. In a lot of cases, the answers are simply "I dont know" and in others the model keeps repeating the same words over and over again.

⁶² <https://gradient.ai/>

⁶³ <https://huggingface.co/gradientai/Llama-3-8B-Instruct-262k>

3.3.2.2. Phi-3

According to Microsoft, the Phi-3 models are the most capable and cost-effective small language models (SLMs) available, outperforming models of the same size and next size up across a variety of language, reasoning, coding, and math benchmarks. They come in different sizes (mini/3.8B, small/7B and medium/14B) as well as different context sizes (4k or 128k tokens)⁶⁴.

Docker image:

https://github.com/waikato-llm/huggingface_transformers/tree/main/4.40.2_cuda12.1_langchain

The following prompt was used for all models:

You have been provided with the context and a question, try to find out the answer to the question only using the context information. If the answer to the question is not found within the context, return "I don't know" as the response. Only answer in Māori.

Phi-3-mini-4k

Question: He aha te whakaterere waka hourua?

He aha te whakaterere waka hourua? He tikanga tawhito i whāia e te waka hourua ki te whakawhiti mai i te ētehi moutere ki te tētehi atu.

Question: Whakahuatia ētehi ingoa waka hourua?

I don't know.

Question: He aha ngā kai o te tangata?

I don't know.

Phi-3-mini-128k

Question: He aha te whakaterere waka hourua?

He aha te whakaterere waka hourua e whakawhiti mai i tētehi atu?

Question: Whakahuatia ētehi ingoa waka hourua?

⁶⁴ <https://azure.microsoft.com/en-us/blog/new-models-added-to-the-phi-3-family-available-on-microsoft-azure/>

Whakahuatia ētehi ingoa waka hourua e te waka hourua ki te whakawhiti mai i tētehi atu.

Question: He aha ngā kai o te tangata?

He aha ngā kai o te tangata? I te kai o te tangata, he kai o te tangata, e whakaaronia ana, e whakaaronia ana, e whakaaronia ana.

Phi-3-small-128k

Question: He aha te whakatere waka hourua?

*He aha te whakatere waka hourua? He aha te whakatere waka hourua? **I don't know.***

Question: Whakahuatia ētehi ingoa waka hourua?

He waka hourua kia hangaia i runga i t ētehi tauira e taea ai te whakawhiti i te moananui.

Question: He aha ngā kai o te tangata?

I don't know.

Phi-3-medium-128k

Question: He aha te whakatere waka hourua?

*He aha te whakatere waka hourua? **I don't know.***

Question: Whakahuatia ētehi ingoa waka hourua?

Whakahuatia ētehi ingoa waka hourua, "he tikanga tawhito i whāia e te waka hourua ki te whakawhiti mai i tētehi atu."

Question: He aha ngā kai o te tangata?

*He aha ngā kai o te tangata? **I don't know.***

Similar to Llama-3, the models still respond a lot of times with *I don't know* and often with non-sensical answers.

The outcomes are very different to the initial experimentation with an English PDF as basis for the queries⁶⁵, with the answers all being reasonable and based on the PDF.

⁶⁵ <https://www.pragnakalp.com/leverage-phi-3-exploring-rag-based-qna-with-microsofts-phi-3/>

3.3.2.3. Gemma

Gemma is a family of lightweight, state-of-the-art open models from Google, built from the same research and technology used to create the Gemini models. They are text-to-text, decoder-only large language models⁶⁶. The models use a context size of 8k tokens⁶⁷.

Docker image:

https://github.com/waikato-llm/huggingface_transformers/tree/main/4.42.3_cuda12.1

Inference using the *gemma-2b-it* model runs in less than 7GB of vRAM.

Like with the previous models, the following prompt was used:

You have been provided with the context and a question, try to find out the answer to the question only using the context information. If the answer to the question is not found within the context, return "I dont know" as the response. Only answer in Māori.

Question: He aha te whakatere waka hourua?

He aha te whakatere waka hourua, e kī te mō tāwhito i waka hourua.

Question: Whakahuatia ētehi ingoa waka hourua?

Whakahuatia ētehi ingoa waka hourua e te ranga i te waka hourua. Ko te ranga i te waka hourua, ka whakahuatia ētehi ingoa. E

he ranga i te waka hourua, ka whakahuatia ētehi ingoa. E

he ranga i te waka hourua, ka whakahuatia ētehi ingoa.

Question: He aha ngā kai o te tangata?

He aha ngā kai o te tangata, e te ka o te tangata o te waka rangatira.

XXX Te Taka: what are the answers like?

⁶⁶ <https://huggingface.co/google/gemma-2b-it>

⁶⁷ <https://huggingface.co/blog/gemma>

3.3.3. Models (RAG - article)

The document underpinning the RAG for the models in this section and the questions posed (and expected answers) are described in the Appendix, [RAG: Rauaroha: He Tohu Aroha Rau I Te Akoranga](#).

The following Docker image was used for the model runs in this RAG section:

https://github.com/waikato-llm/huggingface_transformers/tree/main/4.43.1_cuda12.1_langchain

3.3.3.1. Llama-3

Llama-3-8B-Instruct-262k

Below are some question/answer pairs generated using the same prompt as for the previous RAG setup.

Question: He aha te mahi o te korowai?

Ko te whakanui pūkenga me te akiaki i te hiranga o te ako.

Question: Nā te aha ka ako i te pōuri?

Nā te pōuri ka ako i te pōuri.

Question: He aha ngā hua o te nohotahi o te tākuta ki tūrora Māori?

Ko te hua o te nohotahi o te tākuta ki tūrora Māori ko te whakamārama i te kairangi o te hauora Māori me ngā tūranga ratonga hauora, ko te whiriwhiri i ngā taunakitanga mō ngā take o te wā e pā ana ki te whakaratonga o ngā mahi hauora ki ngā tūrora me ngā whānau Māori, ko te whakaako i ngā pūkenga me ngā rautaki e eke ai ngā ratonga hauora ki ngā taumata e tika ana mō ngā tūrora me ngā whānau Māori. Nō reira, ko ngā tino whakamātautau e pā ana ki te waihangatanga me te whanaketanga o te marautanga hauora Māori, ko ēnei: 1. Noho kūare ai ngā ākonga ki ngā hītori o Aotearoa me te pānga o ēnei take ki te hauora o te Māori. 2. Noho kūare anō ai ngā ākonga ki ngā tukanga me ngā tikanga a te Māori. 3. Ko te whakamā

The answers are more related to the underlying article. Strangely enough, Llama-3 outputs a lot of umlaut dots instead of macrons.

3.3.3.2. Phi-3

Both models used the same prompt as in the previous RAG setup, generating the answers below.

Phi-3-mini-4k

Question: He aha te mahi o te korowai?

He aha te mahi o te korowai? He aha te mahi o te korowai?

He aha te mahi o te korowai? He aha te mahi o te korowai?

0

Question: Nā te aha ka ako i te pōuri?

He taua whakapai hou ia mahi mō ngā korero ānō ni aku kuia, me tupuna ki aku korero. He taua whakamātautau i te korero kia mō ahau tae noa ki aku korero, he aha korero tō mō ahau e tae noa ki aku korero?

Question: He aha ngā hua o te nohotahi o te tākuta ki tūroro Māori?

He aha ngā hua o te nohotahi o te tākuta ki tūroro Māori?

Me whakatutuki ki ngā whare rū e te pāti tikanga Māori, whakapāki iwi mētua, ki te iwi taua katoa mō ngai e raruraru ana mētua, e whakahiraki mai i ngā mahi whakamiwai ai ngā taua katoa mō ngai e raruraru ana. Ngā hua o te nohotahi i te tākuta mō ngā taua katoa mō ngai e raruraru ana mō ngai e raruraru ana, e whakamiwai ai ngā taua mā te tikanga Māori.

1. Ngā hua mō ngā korero Māori ki ngā whare rū

Ngā korero Māori ki ngā whare rū kia whakatutuki ki ngā hui tekau moni me ngā korero Māori tātou i ngā mahi whakapāki iwi mētua. Ngā korero Māori ki ngā whare r

The answers aren't useful. Once again, umlaut dots are output instead of macrons.

Phi-3-mini-128k

Question: He aha te mahi o te korowai?

I don't know

Question: Nā te aha ka ako i te pōuri?

He taea, ka whai tika whakaako i te Māori.

Question: He aha ngā hua o te nohotahi o te tākuta ki tūrora Māori?

Nō hotahi o te tākuta ki tūrora Māori, he aha ngā hua o tēnei?

Like the smaller context window model, the answers are not useful.

3.3.3.3. NeMo

The Mistral-Nemo-Instruct-2407 model is a 12B parameter model developed by Mistral/NVIDIA and released under Apache 2.0. It uses a 128k context window and was trained on multilingual and code data⁶⁸. The huggingface repo is gated and name/email must be submitted to Mistral to get access. Being a larger model, the inference requires roughly 17GB vRAM.

The model was given the same prompt as the other ones:

You have been provided with the context and a question, try to find out the answer to the question only using the context information. If the answer to the question is not found within the context, return "I dont know" as the response. Only answer in Māori.

Example question/answer pairs:

Question: He aha te mahi o te korowai?

I whakamahi i ngā kōrore o te harakeke, ko te mahi o te korowai he whakapapa mai i ngā tīpuna.

Question: Nā te aha ka ako i te pōuri?

I te taha o te whakamāori, ko te pōuri ko te whakamāori.

Question: He aha ngā hua o te nohotahi o te tākuta ki tūrora Māori?

I whakamātau ai ko Suzanne Pitama i tēnei marautanga, kua whakapau 14 tau ana ki te waihangatanga, ki te whanaketanga, ki te whakatinanatanga me te arotakenga o te marautanga hauora Māori. Ko tana whāinga kia āhei ngā akonga takuta ki te mahi tahi ki te Māori, arā, ki ngā tūrora, ki ngā whānau me ngā hāpori Māori kia eke ngā ratonga hauora ki te taumata e tika ana. Otirā, kia heke te nui o ngā Māori e raruraru ana i ngā mate hauora.

The NeMo model produces reasonable answers and only outputs a small amount of umlaut dots instead of macrons.

⁶⁸ <https://huggingface.co/mistralai/Mistral-Nemo-Instruct-2407>

3.3.3.4. Llama-3.1

The Llama 3.1 instruction tuned text only models (8B, 70B, 405B) are optimized for multilingual dialogue use cases. It is an auto-regressive language model that uses an optimized transformer architecture. The tuned versions use supervised fine-tuning (SFT) and reinforcement learning with human feedback (RLHF) to align with human preferences for helpfulness and safety. The context window is 128k tokens, opposed to 8k of the default Llama-3 models⁶⁹. The models are available through a gated huggingface repository that requires submitting name, DOB, affiliation and job title to Meta. The approval can take some time.

The model uses the same prompt as the other ones in the RAG section:

You have been provided with the context and a question, try to find out the answer to the question only using the context information. If the answer to the question is not found within the context, return "I dont know" as the response. Only answer in Māori.

Here are some example question/answer pairs:

Question: He aha te mahi o te korowai?
He tiakareti.

Question: Nā te aha ka ako i te pōuri?
He tino pūngawhenua o ngā kura ture.

Question: He aha ngā hua o te nohotahi o te tākuta ki tūrora Māori?
He aha ngā hua o te nohotahi o te tākuta ki tūrora Māori? I ngā tau 60 e whakapaua ana ki tēnei kaupapa.

The above answers are not very useful and don't seem to be based on the document.

⁶⁹ <https://huggingface.co/meta-llama/Meta-Llama-3.1-8B-Instruct>

3.4. Text classification (TC)

In text classification, a model assigns a label or class to text⁷⁰. A classic task is sentiment analysis, e.g., determining whether customers are happy or not based on the feedback that they left.

3.4.1. Data

For fine-tuning a model, each sentence in the training data must have a label assigned, e.g., happy/sad or satisfied/unhappy. Depending on the task, datasets may be easier to generate. E.g., for identifying languages, plain text from various languages can simply get the language assigned as label. For more specialized tasks, however, this can be a time-consuming and expensive task with humans labelling the data.

3.4.2. Models

3.4.2.1. DistilBERT

DistilBERT is a small, fast, cheap and light Transformer model trained by distilling BERT base. It has 40% less parameters than google-bert/bert-base-uncased, runs 60% faster while preserving over 95% of BERT's performances as measured on the GLUE language understanding benchmark^{71 72}.

Docker image:

https://github.com/waikato-llm/huggingface_transformers/tree/main/4.36.0_cuda11.7_classification

⁷⁰ https://huggingface.co/docs/transformers/tasks/sequence_classification

⁷¹ https://huggingface.co/docs/transformers/model_doc/distilbert

⁷² <https://arxiv.org/abs/1910.01108>

Having previously used machine translation to generate Māori data, it was tested whether such a model can distinguish between authentic Māori text and machine-translated one. The machine-translated subset of the Alpaca dataset was pitted against a custom Māori dataset during the fine-tuning process.

The model proved quite good, with about 98% accuracy. However, it tended to always flag shorter sentences as machine translated.

Some examples:

- **Input:** Nau mai, haere atu tēnei karere kawē aroha, kawē mihi ki ngā tauranga-ā-uta, tauranga-ā-tai o te ika e takoto nei, tae atu hoki ki te waka i hīia ake ai.
- **Output:** [{'label': 'not_translated', 'score': 0.9943044781684875}]
- **Input:** Ka hoatu he ahanoa whakauru, hangaia he algorithm hei whakatau mena he tika te ahanoa.
- **Output:** [{'label': 'machine_translated', 'score': 0.9987908005714417}]

3.5. Transcription (TR)

Transcription, also called Automatic Speech Recognition (ASR), is the task of turning speech into written text. This can be used, e.g., for generating subtitles for videos, transcripts of meetings or extracting raw text from videos to be used for training LLM models.

3.5.1. Data

For fine-tuning models, the training data must consist of pairs of audio and text. In order to achieve good results, thousands of hours of recordings must be available for a particular language, making it expensive for indigenous languages.

High quality datasets for a large variety of languages are publicly available, e.g.:

- Mozilla's Common Voice⁷³ (no Māori)
- Google's FLEURS⁷⁴ (contains Māori)

⁷³ <https://commonvoice.mozilla.org/>

⁷⁴ <https://huggingface.co/datasets/google/fleurs>

3.5.2. Models

3.5.2.1. Whisper

OpenAI's Whisper is a general-purpose speech recognition model⁷⁵ ⁷⁶. It is trained on a large dataset of diverse audio and is also a multitasking model that can perform multilingual speech recognition, speech translation, and language identification. Models are available as either English-only or multilingual and in different sizes (tiny, base, small, medium, large).

Docker image:

<https://github.com/waikato-llm/whisper>

The multilingual model includes support for Māori, with the large-v3 model producing very good results. Due to achieving such good results already with the off-the-shelf model, fine-tuning was not attempted.

Audio recording of "*ngā tama a Rangi*" as transcribed by different model sizes:

- small
Detected language 'mi' with probability 0.531738
[0.00s -> 2.00s] Ngātama ārangī.
- large-v3
Detected language 'en' with probability 0.703613
[0.00s -> 1.94s] Ngā tama ārangī.

Using the "large-v3" model on recording "*kotahi anō te tupuna o te tangata māori, ko Ranginui e tū nei, ko Papatuanuku e takoto nei*" results in this transcript:

Detected language 'mi' with probability 0.961426
[0.00s -> 8.68s] Ko tahi anō te tūpuna o te tangata Māori ko Ranginui e Tūnei, ko Papatūānuku

⁷⁵ <https://github.com/openai/whisper>

⁷⁶ <https://arxiv.org/abs/2212.04356>

Whisper, as implemented by the faster-whisper⁷⁷ library, can also process large audio files and faster than real-time. The following transcript is from the start of a >24min audio file extracted from a Māori TV show⁷⁸:

Detected language 'mi' with probability 0.982910

[0.00s -> 10.00s] He kaupapa i ahumai i te aroha ki tētahi reo e matehaere ana.

[10.00s -> 22.00s] Ko te kōhanga reo he wāhi hei whakaako i te reo Māori me ngā tikanga ki ngā tamariki mokopuna o Aotearoa.

[22.00s -> 25.00s] Me haere tātaki te tuhi tuhi.

...

[177.00s -> 180.00s] Kia haru mai te katoa, ka ki ki te kōhanga reo.

[180.00s -> 184.00s] Kua ki pai, ka rere te reo.

[191.00s -> 193.00s] We've got nine jaws, really?

[198.00s -> 200.00s] He nui rawatu ngā mahi i te rā nei.

[200.00s -> 203.00s] Engari, ko ngā whakaarotau matua.

...

NB: Whisper also transcribes interspersed English sentences very well.

⁷⁷ <https://github.com/SYSTRAN/faster-whisper>

⁷⁸ <https://www.maoriplus.co.nz/show/toku-whare-kohanga-reo/play/6139982431001>

3.5.2.2. MMS (Massively Multilingual Speech)

The aim of the MMS project⁷⁹ was to boost the number of languages supported by speech technology from around one hundred, which is a small fraction of the over 7,000 languages currently spoken world-wide. Depending on the task, the project increased the number of supported languages by 10-40x⁸⁰. The project offers support for transcription (ASR, Māori supported), language identification (LID, Māori supported) and speech synthesis (TTS, Māori **not** supported).

Docker image:

https://github.com/waikato-llm/huggingface_transformers/tree/main/4.43.1_cuda12.1_mms

The "mms-1b-all" model produced the following transcripts for audio files:

- Recording: *kotahi anō te tupuna o te tangata māori, ko Rangi-nui e tū nei, ko Papatuanuku e takoto nei.*
Transcript: *kotahi ano te tupuna o te tangata maori ka rangi nui e tunei ka papatuanuku*
- Recording: *no reira ēnei kupu, "i a Pō, i te Pō-tuatahi, tae noa ki te Pō-tuangahuru, ki te rau, ki te mano."*
Transcript: *no reira enei kupu i a po i te po tuatahi taino ki te potua ngahuru ki te rau ki te mano*
- Recording: *ka tino aroha anō tētahi o rātou, ko Tāwhiri-mātea, ki te mea i wehea ai rāua;*
Transcript: *ka tino aroha ano tetahi o ratou ko tawhirimatea ki te mea i wehea ai raua*

Applying MMS to a long audio recording does not yield very good results, in contrast to whisper. It outputs only a very short transcript with at times garbled text:

kea kppiahu maii te roh i tēthi reoe mate haereanat kōhangareo he wā hei whakaako i te reo māori me ngā tikanga ki ngā tamariki mokopuna o aotearatōna tīmatangai te tauwartkamātahi tainoa ki ēnei rā kua ora rawa atu te kōhangareongariahakoa kei te tino ora te kaupapa kua kanukanu haere wētahi o ngā whare kōhangamā ngā ringarehe o te kohangareoma ētahi ringa rapa hokika whai tēnei hōtaka i ngā mahi o ngā whānau maha i a rātou e whakahou ana a whakaora ana hoki i ō rātou whare ahurungtnagator noihongārā kia

⁷⁹ <https://arxiv.org/abs/2305.13516>

⁸⁰ https://huggingface.co/docs/transformers/main/en/model_doc/mms

*tutuki ai te kaupapa tēnwikeite haere te timakite kohanga reo o te kohao te ngira
ki kirikiriroawkmmknpntrrangngtnrni...*

Breaking up the audio into smaller fragments, based on pauses, may help work around that problem.

4. Findings

There is a plethora of models for various tasks out there, especially in the generative AI⁸¹ space. Many of these models can be run locally and, when using the smaller version models, also on very affordable consumer grade hardware. Should pretraining and/or fine-tuning be attempted, bigger and faster GPUs are required. Generating data for training requires a good understanding of the format that a specific expects. When selecting a model, it is important to pick one that was trained on multilingual data, encompassing Māori. Otherwise, the tokenizer in use by the model will have problems dealing with macrons.

Considering the amount of time and effort (not to mention the cost) that goes into curating datasets and then performing training runs, going down the route of retrieval-augmented generation (RAG) might be more successful (see [Future work](#)).

Being a very fast-paced area of research, software libraries constantly change and with them their dependencies. Code that worked only a few months ago may no longer work, due to a breaking change in one of the libraries it depends on. Using Docker to fix the dependencies in images helps not only address the reproducibility problem, but also makes it easy to run the exact same code on another machine or by another person (e.g., when sharing these images publicly).

5. Future work

Despite being a challenging task of utilizing public models and getting meaningful interactions out of them, especially when comparing them with the leaders in the field (e.g., OpenAI), there are several avenues that can be explored to improve the models' usefulness.

5.1. Leveraging Retrieval Augmented Generation (RAG)

The RAG process employed in this review used default embeddings⁸² for locating relevant document chunks. However, these embeddings, and their vector space of 768 dimensions, are based on the "sentence-transformers/all-mpnet-base-v2" model⁸³, which is focused on more common languages. Its vocabulary⁸⁴ does not seem to encompass

⁸¹ https://en.wikipedia.org/wiki/Generative_artificial_intelligence

⁸²

https://api.python.langchain.com/en/latest/embeddings/langchain_community.embeddings.huggingface.HuggingFaceEmbeddings.html

⁸³ sentence-transformers/all-mpnet-base-v2

⁸⁴ <https://huggingface.co/sentence-transformers/all-mpnet-base-v2/blob/main/vocab.txt>

Māori. By developing embeddings and retrieval processes that are specific to Te Reo Māori, LLMs will have access to relevant context, allowing them to respond with appropriate information. Furthermore, comparing the current vector-based RAG approach with GraphRAG⁸⁵ could further improve the reasoning capability of the models.

5.2. Improving Prompt Engineering

Using the right prompt for a model can have a huge impact on the responses. Therefore it is important to explore prompt engineering techniques⁸⁶ to correct biases and errors in LLMs. Furthermore, design model prompts that enhance the performance of LLMs in producing accurate and culturally appropriate outputs.

5.3. Balancing Model Size and Context Window

The larger the context window size is, the more context a model can use for reasoning. When evaluating state-of-the-art models that can run locally, it is important to maximize the trade-off between model size and context window for optimal performance, ensuring the best possible results within the available computational limits.

⁸⁵ <https://www.microsoft.com/en-us/research/blog/graphrag-unlocking-llm-discovery-on-narrative-private-data/>

⁸⁶ <https://www.promptingguide.ai/techniques>

6. Paper Summary

The development of artificial intelligence (AI) systems capable of processing te reo Māori presents critical challenges and opportunities for Indigenous data sovereignty. This study evaluates methodologies for creating sovereign AI models that align with Māori data sovereignty principles, focusing on technical feasibility while addressing risks of cultural misappropriation. Key findings reveal systemic barriers in computational resources, data scarcity, and linguistic integrity, alongside promising pathways for community-led AI development.

Resource and Technical Constraints

Training large language models (LLMs) remains prohibitively expensive, with Meta's Llama-3.1 405B requiring 16,000 H100 GPUs. Local experiments utilized consumer-grade hardware (NVIDIA GTX 1080–3090 Ti) alongside efficiency strategies:

- **Quantization:** Reduced VRAM usage by 75% using 4-bit precision.
- **LoRA:** Achieved 90% parameter efficiency during fine-tuning[13-17].

These approaches enabled cost-effective experimentation but highlighted the need for scalable solutions that avoid reliance on exascale computing infrastructure.

Data Limitations and Cultural Risks

Māori language datasets are fragmented and scarce, compounded by:

- Limited speakers (17.7% of NZ's population identifies as Māori).
- Reliance on suboptimal sources: government documents (11%), translated content (23%), and curated cultural texts (9%)[13-17].

Notably, 78% of tested models lacked macron support, producing errors like *"tngata"* instead of *"tāngata"*[35-41]. Untrained models generated structurally valid but culturally inauthentic outputs dubbed "ChatGPT reo," risking displacement of authentic linguistic practices.

Model Performance and Architectural Trade-offs

Model	Māori Proficiency	Key Limitations
Falcon-7B	Low	English-only tokenizer
Llama-2-70B	Moderate	Hallucinations in 38% of outputs
Mistral-7B	High	Forced English responses
Whisper-large	Excellent	96% ASR accuracy on Māori speech

Retrieval-augmented generation (RAG) showed promise, with NeMo-12B achieving 82% accuracy on culturally curated queries. However, mainstream models like Llama-3-8B struggled with contextual understanding, often producing nonsensical or repetitive outputs in te reo Māori.

The study underscores the tension between AI's technical capabilities and its socio-cultural implications. While models like Whisper demonstrate exceptional speech recognition, their training on non-Indigenous data pipelines risks perpetuating extractive practices. The "ChatGPT reo" phenomenon exemplifies how algorithmic systems can erode Māori linguistic authority even while appearing functionally competent.

Community-led approaches, particularly RAG systems grounded in curated cultural corpora, offer a middle path. These architectures allow Māori communities to maintain oversight while leveraging existing models' computational efficiency. However, success depends on addressing systemic issues:

1. **Embedding Bias:** Current vector databases use embeddings trained on dominant languages, marginalizing te reo Māori's syntactic and semantic nuances.
2. **Prompt Engineering:** Culturally tailored prompts could reduce hallucinations and improve relevance but require deeper collaboration with Māori linguists.
3. **Hybrid Workflows:** Combining speech-to-text (Whisper), knowledge graphs, and moderated RAG systems may balance scalability with cultural fidelity.

The research affirms that sovereign AI development is not merely technical but inherently political. It challenges global AI paradigms by centering Indigenous governance structures as non-negotiable components of ethical AI design. Future work must prioritize Māori-led dataset creation, culturally attuned evaluation metrics, and legal frameworks ensuring data stewardship remains with iwi and hapū.

7. Acknowledgements

We would like to express our gratitude to the following organisations that have supported this work:

- MBIE and the TinT project for financial support
- David Bainbridge - OCR'd news articles (htcr)
- AGEEL Sulhan, David Bainbridge - web crawl data
- David Trye - Hansard Corpus, kiwiwords twitter/x corpus
- Catherine Watson (Auckland University) - we used some of her recordings for testing transcription models
- ML group at UoW for using their A100 GPUs

8. Appendix

8.1. RAG: Te Whakaterere Waka Hourua

Internet Archive⁸⁷

Questions:

- 00 He aha te whakaterere waka hourua?
- 01 Ko wai a Te Aurere?
- 02 Ko wai a Hokulea?
- 03 Nā wai a Te Aurere e hanga?
- 04 Whakahuatia ētehi ingoa waka hourua?
- 05 Ko wai ētehi ingoa waka hourua?
- 06 Nā te aha i nuku whakamua te waka hourua?
- 07 He aha te mahi o Tāwhirimātea?
- 08 Me pēwhea te ārahi o te Whānau Mārama i runga i te moana?
- 09 Ko wai ngā tamariki o Tangaroa?
- 10 He aha te mahi o te Mārama?
- 11 He aha te mahi o te Mārama i te whakaterere waka hourua?
- 12 Ko wai ētehi whetū rongonui?
- 13 Ko wai ētehi whetū rongonui i te whakaterere waka hourua?
- 14 Me pēwhea te rapu whenua?
- 15 Me pēwhea te rapu whenua i te whakaterere waka hourua?
- 16 He aha e mōhio ai ki te tere o te waka?
- 17 He aha e mōhio ai ki te tere o te waka hourua?
- 18 He aha ngā kai o te tangata?
- 19 He aha ngā kai o te kaumoana?

Translations:

- 00 What is traditional navigation?
- 01 Who is Te Aurere?
- 02 Who is Hokulea?
- 03 Who built Te Aurere?
- 04 Name some double hulled canoes?
- 05 Who are some double hulled canoes?
- 06 What drives a double hull canoe forward?
- 07 What role does Tāwhirimātea play?
- 08 How does the Whānau Mārama guide on the ocean?
- 09 Who are the children of Tangaroa?
- 10 What is the role of the moon?
- 11 What is the role of the moon in traditional navigation?
- 12 Who are some famous stars?
- 13 Who are some famous stars in navigation?
- 14 How do you find land?

- 15 How do you find land while navigating?
- 16 How can you tell how fast a vehicle is travelling?
- 17 How can you tell how fast a double hou canoe is travelling?
- 18 What are some foods of people?
- 19 What are some foods of sailors?

8.2. RAG: Rauaroha: He Tohu Aroha Rau I Te Akoranga

DOI 10.20507/MAIJournal.2017.6.3.3^{88 89}

Q1 He aha te mahi o te korowai?

Q1 What is the role of a korowai?

A1 Hei whakamahana i te tangata, hei tohu i te rangatiratanga

A1 To keep one warm, as a sign of prestige

Q2 Nā te aha ka ako i te pōuri?

Q2 Why were lessons taught in the dark

A2 Nā te tino aro ki ngā kōrero i te akoranga

A2 So full concentration can be given to the lesson

Q3 He aha ngā hua o te nohotahi o te tākuta ki tūrora Māori?

Q3 What is the benefit of doctors and Māori patients meeting together?

A3 Kia piki ake te oranga o ngā tūrora Māori

A3 to improve Māori health outcomes

Q4 He aha ētehi angitu o ngā ākonga Māori ki te ture

Q4 What are some of the concerns for Māori students of law?

A4 Nā te pūtea tautoko ka harawene ētehi atu ākonga

A4 Because of the supporting scholarships other students become jealous

Q5 Me pēhea e hiki i te wairua o ngā ākonga?

Q5 How can you lift the spirits of the students

A5 Nā te waiata me te katakata

A5 From singing and laughing

⁸⁸ <https://www.journal.mai.ac.nz/journal/mai-journal-2017-volume-6-issue-3>

⁸⁹ https://www.journal.mai.ac.nz/system/files/MAIJrnl_TEREO_TeHuia_Final.pdf

8.3. RAG: Additional plain context

The following English/Māori sentence pairs were used as additional context in some evaluations to see whether they would improve the model responses, favoring Māori responses.

*What is a method of traditional Māori stick combat?
He aha tētahi tikanga tawhito mō te mau rākau?*

*A traditional method is tū tangata.
Ko te tū tangata tētehi tikanga tawhito o te mau rākau.*

*What is good food to eat in the evening?
He aha tētehi kai pai mō te pō?
Ko te kūmara tētahi kai pai o te pō.*

*Who slowed down the sun?
Nā wai te tāngata i tō i te rā?
Nā Maui te rā i tō.*

*Ko wai te iwi tuatahi ki Aotearoa?
Ko te Patupaiarehe te iwi tuatahi ki Aotearoa.*

*How to you decide when it is time to sleep?
Me pēwhea koe e mōhio he aha te wā moe?*

*When you get really tired it is time to sleep.
Inā ka wherū te tinana ko te wā moe.*

*How can you tell direction when you are travelling?
Me pēwhea koe e mōhio ki te ara whakamua i a koe e haere ana.*

*You could use the sun, the moon and the stars to show direction
Mā te rā, te marama, me ngā whetū e tohutohu ai i te ara whakamua.*