Low-compute models for low-resource languages: Quantisation of LLMs for English-Māori Translation

Candidate 135

Abstract

Large language models (LLMs) excel in high-resource language tasks, but struggle with low-resource languages. They also have large computational requirements, making them unsuitable for deployment on edge devices. Quantisation offers a way to reduce memory and compute requirements by representing model parameters with fewer bits.

We present the first Māori quantisation study of LLMs. We measure BLEU and chrF translation scores on the FLORES-200 multilingual dataset using Llama-2 13B, a popular open model. We benchmark multiple GGUF settings from unquantised 16-bit down to 2-bit, and find that the unquantised Llama-2 13B fails to produce high-quality translations between English and Māori. We also find that quantisation has a negative impact on BLEU and chrF translation performance, although the impact is larger under human evaluation.

1 Introduction

The application of large language models for low-resource languages is an exciting area, promising to provide natural language processing to underserved groups. However, these models are costly to run locally, and full-precision inference of frontier models is currently only possible on enterprise hardware. We therefore explore the viability of running LLMs for low-resource languages on local hardware, including edge devices. Quantisation offers a solution to this by modifying the bit width of model parameters (referred to as bits per weight/bpw).

This report aims to assess the viability of running LLMs in environments with little memory by benchmarking translation performance on a range of model quantisations. For this purpose, we use translation performance as a proxy to assess general language understanding.

Our research questions are:

1. Does Llama-2-13b-chat produce acceptable English-Māori translations?

2. How does GGUF LLM quantisation affect Llama-2-13b-chat's English-Māori translation performance?

In order to answer our research questions, we have used a range of GGUF-quantised versions of Llama-2 13B, and have benchmarked their English-Māori translation performance on the FLORES-200 dataset. We have also performed a human review of a subset of translations to see if there is a difference between automatic and manual evaluation.

Although model quantisation has been explored for other low-resource language machine translation (Diddee et al., 2022), to our knowledge, it has not been explored for Māori. Therefore, this would be the first study that explores the impact of LLM quantisation on te reo Māori.

Contributions: Our main contributions are:

1. We present the first quantisation study of large language models (LLMs) for English–Māori translation.

2. We benchmark Llama-2-13B-chat from 16-bit to 2-bit, reporting both automatic (BLEU and chrF) and human evaluation scores.

3. We identify a disparity between automatic metrics and human evaluation, demonstrating the limitations of relying solely on automatic metrics for assessing translation quality.

Our experiments show that Llama-2-13B-chat struggles to produce acceptable translations, even without quantisation. We also find that moderate quantisation (e.g., 5-6 bit) introduces only minor degradation in automatic translation metrics (BLEU and chrF). However, our human evaluation reveals more pronounced quality losses that BLEU and chrF do not detect, especially at low bit depths.

This suggests that automatic metrics may underestimate the real-world impact of quantisation on translation quality.

Report structure: In this report, we review related prior work and discuss its relevance to our study. We provide an overview of the benchmark dataset used, explain our methodology (including details of the model, quantisation approach, and evaluation procedures) and present our results accompanied by an error analysis. Lastly, we discuss limitations and ethical considerations before concluding.

2 Related Work

Robinson et al. (2023) benchmarked gpt-3.5-turbo on FLORES-200, finding that it is competitive with traditional MT systems for high-resource languages, but lags behind for low-resource languages including Māori. This shows that even state-of-the-art LLMs struggle with low-resource translation, and supports our decision to focus on Māori as a case study.

Similar model: Fossen (2024) benchmarked a selection of frontier models on FLORES-200. While GPT-4 and Google Translate achieved average scores of 46.3 and 46.6 chrF when translating Māori, Llama-2-70b-base lagged behind with an average score of 29.6. For this project, we are using the smaller Llama-2-13b-chat model because of our resource limitations. This choice should result in lower performance. In addition, our approach uses a zero-shot prompt instead of five-shot. This can produce less accurate translations (Peng et al., 2023). However, we are using the chat model as opposed to the base model, which may improve translation performance because of its improved ability to follow tasks.

LLM MT Quantisation: Jin et al. (2024) evaluated 4-bit quantised LLMs across many benchmarks including FLORES-200, and reported that they retain performance comparable to full-precision models. This suggests that we may see performance being mostly conserved across most quantisation levels. However, they do not report specific scores for Māori.

Alternatives to Quantisation: Instead of quantisation, an alternative technique used for compressing models is distillation, which trains a smaller model on the outputs of a larger one. However, Diddee et al. (2022) found that distillation is a ”brittle” technique, meaning that performance varies heavily depending on priors like the amount of synthetic data, student architecture, training hyperparameters and teacher model confidence. In contrast, they find that quantisation offers a more robust method of model compression, especially for low-resource languages. Because of the scope and resources available for this project, we have not explored distillation and focused solely on quantisation.

Automatic vs Human Evaluation: Marchisio et al. (2024) benchmarked multilingual LLMs on the FLORES-200 dataset as well as performing a human evaluation, and found that automatic metrics such as BLEU severely underestimate the quality loss when applying quantisation, noting that “a 1.7% average drop in Japanese across automatic tasks corresponds to a 16.0% drop reported by human evaluators”. This can imply that small changes in our BLEU or chrF correspond to large changes for our human evaluation.

To summarise, there exists substantial work on the effect of quantisation on translation performance as well as LLM translation performance on Māori. However, to our knowledge, there has not been a study that specifically combines the two. If so, this would be the first study that investigates how LLMs process the Māori language when quantised.

3 Dataset

The dataset used for this project is the FLORES-200 dataset (Team et al., 2022), which provides multilingual data for 204 languages, including many low-resource languages.

Authors & Provenance: The dataset was originally created by Meta’s FAIR team under their No Language Left Behind project (Team et al., 2022), now being maintained by the Open Language Data Initiative (OLDI). For this project, the original Meta dataset has been used.

The dataset is sourced from Wikimedia projects such as Wikinews, Wikijunior and Wikivoyage. The test set contains 1012 sentences, which has previously been shown to be adequate for reliable machine translation metrics (Koehn, 2004).

Stats & File Structure: The dataset has an average of 23.18 words and 1.11 sentences per line (entire dataset). The average words per line is 32.50

for Māori and 24.10 for English. The average sentences per line is 1.12 for both Māori and English.

The dataset contains two folders, one dev set and one devtest set (we have used the devtest set for our experiment). Both sets contain files for every language, with each line containing a given sentence. Bitext can be created for any language pair by extracting the same lines across two files. Since every sentence in the dataset has been manually translated and verified in every language, we have gold standard data. This allows for automatic metrics to be calculated with BLEU and chrF.

Concerns: Although the FLORES-200 dataset is a highly valuable resource, concerns have been raised regarding its provenance, linguistic diversity, and language coverage. We address these issues in detail in the Ethical Considerations section.

4 Methods & Experiments

Model, Quantisation Approach: All models were fetched from the Hugging Face hub. The baseline unquantised model was meta-llama/Llama-2-13b-chat. The quantised models used were fetched from TheBloke/Llama-2-13B-chat-GGUF. All quantisations from the repo were benchmarked. We chose GGUF quantisation over AWQ because of its comparatively wider support as of 2025, and over GPTQ because of its flexible CPU/GPU offloading capability, making it a realistic choice for current edge computing use cases.

Alongside the FP16 unquantised model with 16bpw, the GGUF quantisations tested are as follows:

`Q8_0`, `Q6_K`, `Q5_K_M`, `Q5_K_S`, `Q5_0`, `Q4_K_M`, `Q4_K_S`, `Q4_0`, `Q3_K_L`, `Q3_K_M`, `Q3_K_S`, `Q2_K`

GGUF filename Explanation: For GGUF filenames, the Q flag signals the quantisation level, from 8bpw down to 2bpw. The type-0 approach compresses each 32-weight block of parameters to signed integers as well as a single scale. The type-K approach allows for more specificity at a minor memory cost, and arranges 128–256-weight super-blocks that carry extra 4–8-bit scale metadata for every sub-block. For type-K quantisation, the S/M/L flags denote mix recipes: S keeps every tensor at the stated bit-width, M upgrades a subset of error-sensitive matrices by +1 bit, and L upgrades even more tensors by +2 bits. In summary, we can expect lower quantisation levels to yield higher accuracy. We can also state that type-0 quantisation is more naive and straightforward than type-K. For K-quantisation, we can expect that the L flag should give more accurate results than M, which in turn should give more accurate results than S. However, since our main interest is the quantisation level instead of specific quantisation approaches, we will often average different approaches at the same quantisation level when discussing our results.

Prompt: The following zero-shot prompt was used on all models:

Zero-Shot Translation Prompt

```
Translate the following from {source_name
} to {target_name}. Only write one
additional line with nothing extra:

Source ({source_name}): {source_text}
Translation ({target_name}):
```

Several zero-shot prompts were evaluated to identify the one that most closely adhered to the task instructions; however, no comprehensive prompt engineering process was undertaken.

Hyperparameters: For inference, the baseline model was loaded using the `transformers` library (Wolf et al., 2020). GGUF-quantised models were loaded using the `ctransformers` library (Marella, 2023). When loading these models, we left model hyperparameters at defaults, assuming that the two libraries use similar defaults. This oversight emerged post-experiment: the baseline model was accidentally initialized with different sampling hyperparameters (unquantised: temperature = 0.6, top-p = 0.6; quantised: temperature = 0.8, top-p = 0.95). The ramifications of this mismatch are examined in the Limitations section.

Pre-processing: No pre-processing was applied to the dataset before conducting the experiment. However, when preprocessing outputs for evaluation, only the first line of outputs was used, since following lines usually contained superfluous clarifying statements. The text was tokenised by converting to lower case, removing punctuation, and splitting on whitespace. Māori macron characters were not normalised. BLEU and chrF were computed with `nltk`'s `corpus_bleu` and `corpus_chrf`. These scores were multiplied by

100 to produce conventional BLEU and chrF percentage scores.

Evaluation: We chose BLEU and chrF for evaluation, as they are among the most widely used metrics in machine translation research. Their widespread adoption makes them suitable for comparing our results with prior work.

BLEU (Bilingual Evaluation Understudy) (Papineni et al., 2002) is the most widely used metric for translation performance. It measures word-level n-gram overlap precision between a candidate translation and one or more reference translations under the assumption that higher n-gram precision correlates with better translation quality. Since this can allow overly short sentences with high overlap to gain high scores, a brevity penalty is applied.

ChrF (Character n-gram F-score) (Popović, 2015) is another widely used metric for translation performance. Instead of word-level n-grams, chrF compares character-level n-grams. Being an F-score, it accounts for both precision and recall.

In addition to quantitative metrics, we have included human evaluation from a Māori native speaker. The native speaker provided ratings ranging from 1 to 10 as well as insights on Māori orthography. This evaluation is based on a small selection of translations: We randomly sampled six sentences, then picked translations from three precision levels: The unquantised baseline, the Q4_0 model, and the Q2_K model. Each sentence was translated in both directions, resulting in 36 manual ratings.

Code & Sources: The codebase used in this project was repurposed from a previous project measuring translation performance on frontier API-only models (Fossen, 2024). Software development was aided by ChatGPT (OpenAI, 2024) and Microsoft Copilot (GitHub, 2024) for tasks including code refactoring, generating plots, searching documentation, exploratory data analysis, and crash debugging. The project code is available on GitHub (Fossen, 2025).

5 Results

5.1 Metric-Based Analysis

Unquantised Performance: Our baseline unquantised FP16 model achieved an average BLEU score of **3.35** (**2.11** BLEU eng→mri, **4.59** BLEU mri→eng). For chrF, it achieved an average of **25.91** (**22.42** eng→mri, **29.4** mri→eng) (Figure 1). These are low scores and suggest that the translations are not of acceptable quality. It is worth noting that performance is lower when translating into the lower-resource language. This is a pattern across all quantisation settings and is consistent with previous work (Fossen, 2024).

Quantised Performance: Moving to 8, 6 and 5-bit quantisation, BLEU scores dropped by **8.7%** on average, while chrF scores actually *improved* by **1.2%**. This suggests that for higher quantisation levels, performance loss is minimal.

Performance degradation is more pronounced for 4-bit quantisation, with a **19.7%** and **5.2%** drop for BLEU and chrF, respectively. Contrary to expectations, the naive type-0 quantisation performs better than type-K, and even within type-K, more frugal mixes (S and M) perform better.

At 2-bit quantisation, BLEU scores are significantly lower, showing a drop of **23.2%** compared to the baseline. chrF scores, however, remain comparatively stable, with the 2-bit quantisation showing only a **4.4%** drop from the baseline.

The average performance cost per bit of quantisation (bpw) was **0.044 BLEU** and **0.052 chrF**, showing that quantisation has a small impact on translation performance when using automatic metrics.

In summary, our metric-based analysis suggests that translation performance is already low for the full-precision model. Furthermore, quantisation up to 4-5 bits does not further meaningfully degrade performance, while further quantisation leads to significant quality loss.

5.2 Manual Evaluation

In our manual evaluation, we observed that although most of the translations were meaningful and grammatically correct, they did not carry the same meaning as the original text, rendering them unusable.

Incorrect Māori Orthography: Although our automatic BLEU and chrF metrics show little further degradation in quality, the degradation is more apparent when inspecting translations manually. For example, while the unquantised and 4-bit quantised models produced reasonable words that fit within Māori orthography, the 2-bit quantisation produced words that cannot exist in the Māori language due to use of characters that are not present in the Māori alphabet (e.g. using ”Q” in the hallucinated word ”teteqwe”). The less precise models

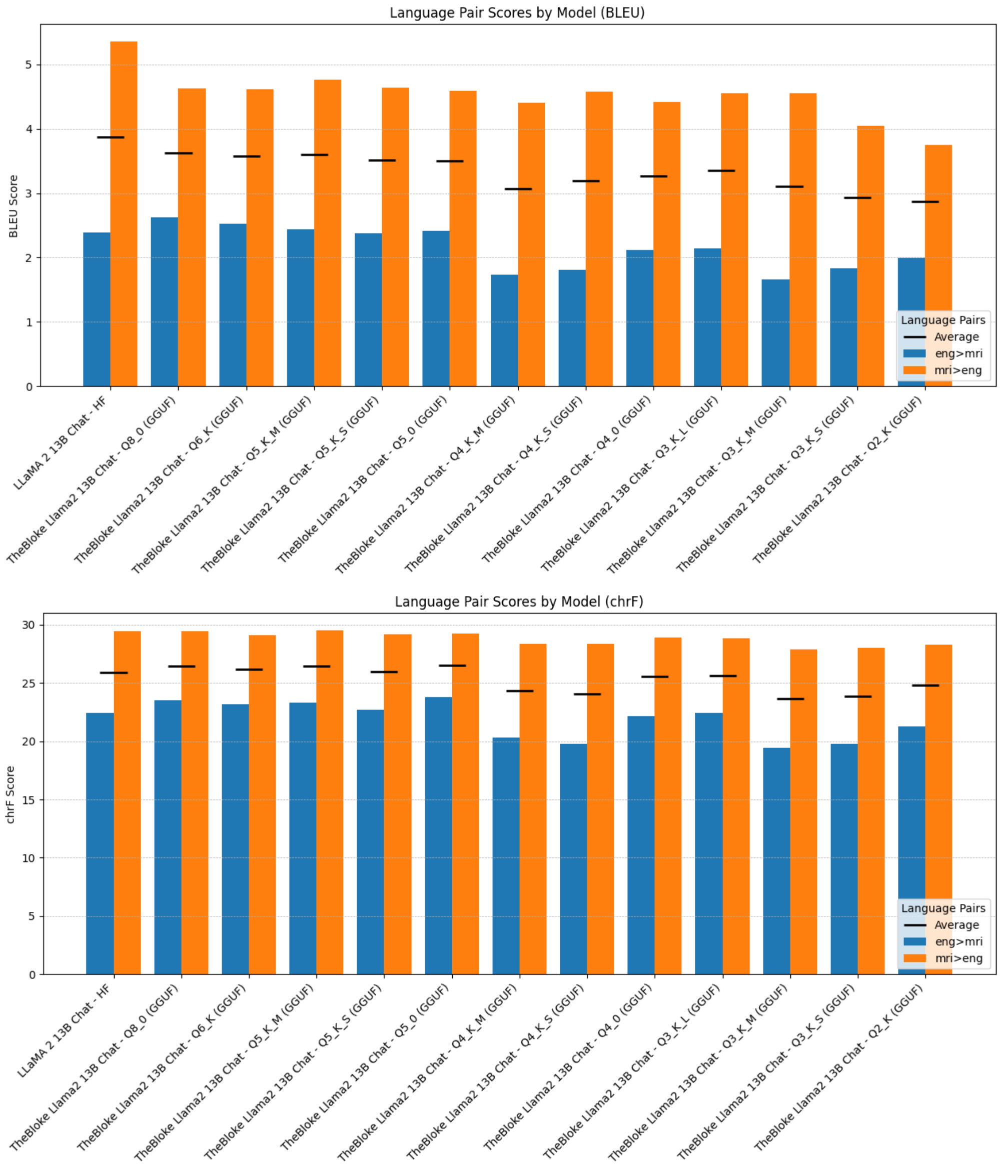

Figure 1: Translation performance between English and Māori of Llama-2-13B-chat under different quantisation levels. Scores for both translation directions are shown, and horizontal lines represent averages. Both scores are shown as percentages.

also frequently perform transliteration or English words, e.g. ”effect”→”efekti”, even though Māori words do not use consonant clusters like ”kt”.

Comparison to Metric-Based Analysis: The Quantitative results of our manual evaluation support these qualitative observations. We found Q4_0 translations were manually rated **23%** lower then full-precision translations, at **3.08** compared to **4.0** out of 10. The Q2_K translations were lower still at **2.50 out of 10**; approximately **38%** lower.

This **38%** drop is more significant than the ones observed using automatic translation metrics (**-23.2%** for BLEU and **4.4%** for chrF). This suggests that automatic metrics underestimate performance loss compared to manual metrics.

Cultural Tropes: One notable qualitative observation was a strong thematic bias towards concepts related to ancestors, tradition and nature. A manual tally of the selected translations found that **53%** of them contained hallucinations with overly poetic language revolving around cultural tropes. Certain terms, such as ”whakapapa” (genealogy) and whenua (land) appeared in English-to-Māori translations with disproportionately high frequency. A simple search in LLM outputs found that while **17%** (2199 in total) of the LLM translations used the word ”whakapapa”, *none* of the reference translations did. This may indicate that translation has an essential cultural component which should be taken into consideration, although that consideration remains outside the scope of this paper.

> Cultural Tropes in Mistranslation
>
> **Source Text (Māori):**
> *Kei te whakakotahi te Ipurangi i ngā pūmotu o te papatipu me te whakawhitinga whaiaro anō hoki.*
>
> **Reference Translation (English):**
> *The Internet combines elements of both mass and interpersonal communication.*
>
> **Q4_0-quantised Model Output:**
> *The wind howls and moans as it rips through the papatu (sedge) and carries away the whispers of the ancestors.*

In the example above, the model output demonstrates a clear failure in translation from Māori to English. The model hallucinates a poetic, unrelated sentence in English.

In summary, our manual evaluation coincides with our metric-based analysis regarding RQ1: Translation performance is unacceptably low, even for the unquantised model. However, for RQ2, our manual evaluation finds an even larger disparity between quantisation levels compared to our automatic metrics.

6 Error Analysis

From both a metric-based and manual evaluation perspective, Llama-2-13b-chat performed poorly regardless of quantisation setting, and can not be used reliably for machine translation. However, metric-based and manual evaluation differs more when attempting to answer the second research question: BLEU and chrF scores seem to underestimate the impact of quantisation compared to manual evaluation, consistent with Marchisio et al.'s findings (2024).

One interesting finding is the bias toward cultural tropes. The reason for this might be due to the examples of Māori text in the model's training data reflecting this. Another explanation may be due to how Māori language and culture is described in English and other languages, which could present a stereotypical image of the Māori language as revolving around these themes. However, because of Meta's failure to disclose their training data, exploring these hypotheses remains unduly difficult.

One unexpected outcome was that some quantisation settings performed better than their higher-precision counterparts. This is most evident with respect to chrF scores, where some quantisation settings outperform the base model. This is consistent with Marchisio et al. (2024) who also find that quantisation can occasionally bring benefits. However, it is unknown whether our result is statistically significant or if it is due to randomness.

A persistent problem was that the model would fail to follow translation instructions, often responding with additional lines and clarifying statements. To circumvent this, we only considered the first line of the model's output when performing automatic and manual evaluation. However, this will cause issues if the model outputted a clarifying statement before the actual translation.

We initially tried to use the base 13B model instead of the chat model to perform translations. However, this model was difficult to prompt sucessfully, even when using a five-shot prompting approach. When we switched to the chat model, prompting the model to follow instructions became substantially easier, and so a zero-shot prompt was used for simplicity.

7 Limitations

Incorrect Hyperparameters: A major limitation of this study is that the hyperparameters for the base model were different to the quantised models, which undermines a direct comparison. Unfortunately, time and resource constraints did not allow for a second run of the experiment with fixed hyperparameters. Peng et al. (2023) noted that higher temperature values can lead to less accurate translations. This suggests that the unquantised model (which had a lower temperature at 0.6 compared to 0.8) may have performed more similarly to the quantised models than these figures suggest. Given more time, we would have performed a full re-run of the experiment to ensure comparable results. For future work, it remains crucial to ensure that model hyperparameters are identical across runs.

Small Model: Since we used a relatively small 13B model, the translation performance was very low. If a larger model had been used, the performance would have been substantially higher, and the impact of quantisation could also have been different. Because of resource and time constraints, we did not benchmark a larger model, although this is a natural direction for future work.

As mentioned, the misconfigured hyperparameters present a major flaw in this work. In the future, a more rigorous experimental setup would be required. Other areas for improvement include using state-of-the-art models to see if model size affects quantisation impact, employing a systematic prompt engineering process to get the most out of the models, and a larger user study to get more definitive manual evaluation metrics.

8 Ethical Considerations

Disclaimer: The safety considerations discussed in this section are framed in the context of a potential scaled-up version of this project where users would be interacting with the model. Being a small-scale experiment without external users, this current study presents little ethical risk. Still, the results presented in this project may influence others' perception of this model, making it important to address these ethical considerations.

Dataset Concerns: The FLORES-200 dataset used for evaluation was created by Meta. The original paper states that a high standard was applied to human translators, but details regarding compensation are absent (Team et al., 2022). Furthermore, it remains unclear whether the dataset includes a diverse range of Māori dialects and regional language variations. This makes it difficult to know whether the language models are able to capture a diverse range of te reo Māori dialects. This can in turn be important to ensure representation across Māori iwi/tribes and to avoid marginalising certain linguistic variations.

Cultural Bias: Large language models are known to generate biased outputs, especially if the training data contains culturally insensitive or inaccurate information. Language preservation and accuracy are sensitive issues in the context of low-resource languages. Because of this, the development of any NLP system for low-resource languages should be aware of these biases and make an effort to align with the language's linguistic and cultural expectations. If this is not considered, misalignment could result in outputs that distort cultural nuances.

Security issues: Using LLMs for any task, including machine translation, inherently carries jailbreaking risks common to LLMs. The model could be manipulated into performing tasks other than translation, as well as generating toxic, harmful, misleading, or culturally insensitive content.

Another security concern is the potential exposure of sensitive information when using the model. Although running models locally can reduce this risk, it does not eliminate it entirely. When models are deployed by companies or third-party providers, users may need to transmit sensitive data over the internet, which reintroduces the risk of exposure.

Language Misrepresentation: Although this system is intended to benefit language preservation efforts and Māori-speaking communities, misrepresentations of the language can hamper Māori revitalisation efforts by propagating incorrect linguistic or cultural information. Inaccuracies in translation carries a risk of reinforcing stereotypes or misinformation. Quantisation evidently lowers translation quality, which can exacerbate this issue.

9 Conclusion

In this report, we explored the impact of quantisation on the translation performance of large language models for English-Māori translation. We benchmarked Llama-2-13B-chat at multiple GGUF quantisation levels from full-precision 16-bit down to 2-bit. Translation performance was evaluated using both automatic metrics (BLEU, chrF) and human evaluation.

We found that the baseline, unquantised 16-bit Llama-2-13B-chat model already exhibits inadequate English-Māori translation quality. Quantisation from 8-bit down to around 5-bit led to only minor degradation according to automatic metrics. However, our manual evaluation revealed more significant quality loss at lower quantisation levels, suggesting that BLEU and chrF underestimate the actual deterioration of translation quality. Furthermore, qualitative analysis revealed that the translations frequently referenced cultural tropes and produced hallucinations that were not related to the source text, which poses a substantial risk of language misrepresentation.

Addressing our research questions explicitly, we conclude that Llama-2-13B-chat fails to produce acceptable English-Māori translations. Furthermore, GGUF quantisation minimally impacts translation quality according to automatic metrics at moderate precision levels, but significantly reduces translation quality according to human evaluation.

Several limitations were identified in our study, most notably the inconsistent hyperparameter initialisation between the baseline and quantised models and the limited amount of manual evaluations. Future research should investigate larger, state-of-the-art models, ensure consistent hyperparameter initialisation expanding manual evaluations, and systematic prompt engineering.

We hope these results contribute to future efforts in improving large language models for low-resource languages, as they show the need for further development in NLP for low-resource languages and for the use of human evaluation alongside automatic metrics.

References

Harshita Diddee, Sandipan Dandapat, Monojit Choudhury, Tanuja Ganu, and Kalika Bali. 2022. Too brittle to touch: Comparing the stability of quantization and distillation towards developing lightweight low-resource mt models.

Luca Fossen. 2024. Large language models for low-resource languages. Technical report, AI Institute at Waikato University. Report, Available at https://lucafossen.me/pub/fossen-llms-for-lrls-2024.pdf.

Luca Fossen. 2025. llm-translation-testing. A test harness for numerically evaluating machine translation performance on a range of large language models. https://github.com/lucafossen/llm-translation-testing.

Microsoft GitHub, OpenAI. 2024. Github copilot. Software used for code refactoring, documentation search, data analysis, and debugging. https://github.com/features/copilot.

Renren Jin, Jiangcun Du, Wuwei Huang, Wei Liu, Jian Luan, Bin Wang, and Deyi Xiong. 2024. A comprehensive evaluation of quantization strategies for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12186–12215, Bangkok, Thailand. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Kelly Marchisio, Saurabh Dash, Hongyu Chen, Dennis Aumiller, Ahmet Üstün, Sara Hooker, and Sebastian Ruder. 2024. How does quantization affect multilingual LLMs? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15928–15947, Miami, Florida, USA. Association for Computational Linguistics.

Marella. 2023. ctransformers: Python bindings for transformer models implemented in c/c++ using ggml library. https://github.com/marella/ctransformers. Accessed: 2025-05-14.

OpenAI. 2024. Chatgpt (may 14 version). Software used for code refactoring, documentation search, data analysis, and debugging. https://chat.openai.com.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318.

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards making the most of ChatGPT for machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5622–5633, Singapore. Association for Computational Linguistics.

Maja Popović. 2015. chrf: character n-gram f-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation (WMT)*, pages 392–395.

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for high- (but not low-) resource languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics.