Evaluating the LLM and NMT Models in Translating Low-Resourced Languages

Julita Pucinskaite
Lancaster University, United Kingdom
julitapucinskaite@gmail.com

Ruslan Mitkov
University of Alicante, Spain
r.mitkov@lancaster.ac.uk

Abstract

Machine translation has significantly advanced due to the development of transformer architecture, which is utilised by many modern deep-learning models. However, low-resource languages, such as Lithuanian, still face challenges stemming from the limited availability of training data and resource constraints. This study examines the translation capabilities of Neural Machine Translation (NMT) models and Large Language Models (LLMs), comparing their performance in low-resource translation tasks. Furthermore, it assesses the impact of parameter scaling and fine-tuning on their effectiveness in enhancing model performance. The evaluation showed that while LLMs demonstrated proficiency in low-resource translation, their results were lower compared to NMT models, which remained consistent across smaller variants. However, as model size increased, the lead was not as prominent, correlating with automatic and human evaluations. The effort to enhance translation accuracy through fine-tuning proved to be an effective strategy, demonstrating improvements in vocabulary expansion and structural coherence in both architectures. These findings highlight the importance of diverse datasets, comprehensive model design, and fine-tuning techniques in addressing the challenges of low-resourced language translation. This project, one of the first studies to focus on the low-resourced Lithuanian language, aims to contribute to the broader discourse and ongoing efforts to enhance accessibility and inclusivity in Natural Language Processing.

1 Introduction

The field of Natural Language Processing (NLP) has been essential in enhancing access to information and promoting inclusivity across different languages. Machine Translation (MT) was developed to utilise computers in overcoming communication gaps and facilitating cross-linguistic cooperation, with early efforts focusing on translating Russian to English. However, despite significant advancements in LLMs, MT and NLP in general, many low-resourced languages remain underrepresented and overlooked by the rapidly growing AI industry.

It is worth noting that Machine Translation has long been a key focus in NLP with the aim of enabling computers to translate natural language automatically. Initially, the field was dominated by the Rule-Based (RB) approach, which relied on manually constructed linguistic rules and dictionaries. However, this method was prone to error, resource intensive and had scalability implications when transferring rules between different languages (Wang et al., 2022). Due to these limitations, interest in RB systems declined, leading to a slowdown in the progress within the MT field. Nevertheless, some continued, resulting in the development of highly accurate RB systems such as Systran and DeepL, while they later transitioned first to statistical and after that to neural network-based architectures.

The field saw meaningful breakthroughs with the adoption of corpus-based methods following the Statistical Machine Translation (SMT), which was reintroduced in the early 1990s by IBM researchers (Brown et al., 1990). SMT leverages large parallel texts and probabilistic models to make predictions on the most likely translation. Initially, these systems relied on single-word mappings, although this introduced many errors in semantic meaning and word reordering, leading to a shift toward phase-based translation (Lopez,

Proceedings of the First Workshop on Comparative Performance Evaluation: From Rules to Language Models associated with RANLP 2025, pages 123–133, Varna, Bulgaria, Sep 11, 2025.
https://doi.org/10.26615/978-954-452-102-8-013

2008). This approach was the foundation for an early version of the Google Translate engine. Despite these advancements, SMT struggled with long-distance word ordering and data sparsity issues, particularly for linguistically distant language pairs (Wang et al., 2022).

The introduction of deep learning techniques, such as a sequence-to-sequence model structure, transformed MT. These models, powered by neural networks utilised an encoder-decoder framework that mapped input sentences to variable-length vector representations, ensuring the retention of sentence structure and meaning (Sutskever, 2014). The addition of the attention mechanisms allowed the decoder layer to focus solely on the relevant input encodings, improving translation fluency and overcoming SMT weaknesses (Bahdanau, 2014). This neural process was extended to multilingual machine translation, where shared encoded representations could be supported by multiple decoder layers for different target languages (Dong et al, 2015).

The Transformer model architecture revolutionised NLP by introducing self-attention mechanisms, removing the need to use recurrence and process one token at a time. Unlike earlier models, these properties allow the model to read all tokens simultaneously, capturing broad contextual relationships regardless of sentence length. This parallel processing led to a significantly faster training on large datasets, making Transformers the foundation of NMT models and LLMs (Vaswani, 2017). NMT models follow a sequence-to-sequence framework, mapping an input sequence from the source language to the target language. Where LLMs are typically categorised as auto-encoding or auto-regressive models, either using encoder or decoder-only architectures, with the latter being more frequent and following the objective of accurately predicting the next token in the sequence (Dong et al., 2019).

The performance of LLM and NMT models is dependent on the availability and quality of the training corpus. These models typically rely on high-resourced languages, such as English and German, with low-resourced languages receiving significantly less representation due to data limitations (Scao et al., 2022). NMT models are usually pre-trained on parallel corpora, which enables a comprehensive representation of language distribution. In contrast, LLMs are trained on diverse texts without targeting multilingualism, which often limits their ability to support low-resource tasks (Paupard, 2024).

To address this disparity, researchers reinforced insufficient parallel corpora with monolingual data (Znang and Zong, 2016). However, a high monolingual data ratio can diminish models learning outcomes, calling for back-translation, which automatically incorporates translations to monolingual texts (Sennrich et al., 2015).

An equally critical aspect is dataset quality, particularly in low-resource settings. A study found that noisy texts can drastically degrade translation accuracy, making cleaned and filtered datasets essential for reliable training (Khayrallah and Koehn et al., 2018). This is highly relevant for underrepresented languages, where datasets are often accumulated using web scraping techniques such as Common Crawl, which collect texts from various internet sources (Toral et al., 2017; Baack, 2024). These findings highlight the key challenges for both NMT models and LLMs that require high volumes of training data but are constrained to limited, low-quality, low-resourced language texts.

The open source and distillation techniques seek to bridge this gap and support a transparent and community-driven development process to direct a more inclusive and comprehensive language technology (White et al., 2024). While advocates for the closed-source design argue that it offers better security and data protection guarantees (Xi, 2025). Despite these claims, closed-source models remain vulnerable to various security risks, including adversarial attacks, suggesting that their motivations may be ineffective (Das et al., 2025).

To utilise both open-source accessibility and closed-source performance, researchers have turned to knowledge distillation, where smaller student models learn from larger teacher models. This technique reduces computational demands while ensuring high accuracy and maintainability of core capabilities (Hsieh et al., 2023). The effectiveness was demonstrated by models like Deepseek, which outperformed state-of-the-art models in multiple evaluation benchmarks (Deepseek-AI, 2025).

The present study will experiment with the distilled versions of both NMT and LLM, including NLLB and Gemma models (Costa-Jussà et al., 2022; Team et al., 2024). Although NLLB is fully open-sourced, Gemma follows an open-weights approach where only the model's parameters are made available without source

code. While not as transparent as open-source, this still enables customisation and adaptation, supporting resource-constrained teams working on low-resourced language tasks (Zhao et al., 2023).

Finally, it is worth noting that the field of Low-Resource NLP has gained significant attention in recent years, as demonstrated by the growing number of research contributions addressing data scarcity and model adaptability challenges, further emphasising the need to improve machine translation for languages like Lithuanian (Pakray, 2025). The present study is significant in highlighting the insufficient support and inclusion of Lithuanian, a low-resourced language, in modern deep-learning tools and LLMs, an area of study that has received little attention. Researchers in developing translation models often neglect the underrepresented languages due to the limited availability of parallel corpora, which are essential for training accurate translation systems (Chakravarthi et al., 2019). As a result, models trained on small or insufficiently diverse datasets often produce inaccurate translations and hallucinations (Poupard, 2024).

The findings from this study aim to contribute to the enhancement of translation technology, making NLP tools more inclusive and accessible for speakers of less commonly spoken languages. Furthermore, by understanding the limitations of pre-trained models and the benefits of fine-tuning, this research can provide insights in directing future machine translation efforts for other low-resource languages.

The rest of the paper is structured as follows: Section 2 outlines related work, Section 3 presents the methodology and Section 4 provides evaluation results, discussion and error analysis. Finally, Section 5 summarises the work with a conclusion.

2 Related Work

Several studies have addressed the disparity in lesser-spoken languages by developing more linguistically inclusive models. The No Language Left Behind (NLLB) team supported over 200 underrepresented languages by training a multilingual NMT model on high-quality parallel and monolingual datasets and adopting self-supervised learning. These unconventional training methods demonstrated enhanced translation performance in low-resource settings even for languages where explicit training was not undertaken (Costa-Jussà et al., 2022).

Martins et al. (2024) trained the EuroLLM model to address the lack of open-weight LLMs for European languages. The authors used a parallel corpus that included nearly an equal number of English and non-English representations. Their findings indicated that carefully curated datasets and a custom tokeniser enabled the model to outperform much larger competitors in translation tasks.

Nakvosas et al. (2024) discussed the insufficient number of Lithuanian language tokens in the Llama model. They employed a supervised fine-tuning (SFT) technique to improve the model's performance in English-Lithuanian tasks. This approach involved training a pre-existing model on a high-quality custom dataset, allowing it to enhance its learning and generation accuracy, especially in handling previously unseen data (Church et al., 2021).

Another fundamental challenge is the high computational costs associated with training LLMs and NMT models as they often contain billions of parameters, requiring extensive memory, storage and processing power (Hadi et al., 2023). These demands create significant barriers for smaller research teams, especially in underrepresented linguistic communities.

To address resource constraints, researchers have explored performance-efficient fine-tuning (PEFT) techniques. One widely adopted approach is quantisation, which reduces the precision of model parameters (e.g., to 8-bit or 4-bit), lowering memory usage without experiencing major performance loss (Dettmers et al., 2024). Low-Rank Adaptation (LoRA) further optimises resource requirements by applying fine-tuning only to targeted layers, preserving strong multilingual performance while reducing trainable parameters (Hu et al., 2021). These techniques provide effective solutions to optimise resource usage, democratising access to LLMs and NMT models for low-resourced language researchers.

Finally, a key research question is whether LLMs can match or surpass well-established NMT models in low-resource language translation. While multilingual LLMs such as Gemma and Llama have demonstrated effectiveness in high-resourced translation tasks, achievements in low-resourced languages, such as Lithuanian, have often remained undiscovered. Furthermore, LLM architecture may suffer from accuracy loss and hallucinations, where models generate fabricated

information when handling large multilingual datasets (Dong, 2024).

Further research is essential to assess the genuine performance of LLMs on underrepresented languages and to determine the trade-offs between model size and fine-tuning in translation quality, thereby contributing to more inclusive NLP systems.

3 Methodology

The adopted methodology, detailed in this section seeks to reply to the following questions:

1. How do pre-trained LLMs and NMT models perform in low-resourced language translation?
2. Does fine-tuning improve translation accuracy? Is it comparable to parameter scaling?

In particular, we outline the data, models and evaluation methods employed in this study and acknowledge experimental limitations.

3.1 Research Design and Data Collection

This study follows an empirical, quantitative approach to evaluate model performance. Models, datasets and fine-tuning tools were obtained through the Transformers library, which provides open-access NLP resources.

Supervised Fine-tuning (SFT) requires rich translation examples. Although scaling laws suggest that the optimal dataset size should be proportional to the model's parameter number. For example, a 1.3 billion parameter model (NLLB) would need around 30 million sentences (Hoffman et al., 2022). However, further recent research shows that smaller, diverse datasets can still yield sufficient performance (Oliver and Wang, 2024; Zoph et al., 2022).

Given resource limitations and limited text availability, a dataset of 300,000 English-Lithuanian sentence pairs was compiled from the following corpora:

Medical Corpus – domain-specific translations with complex terminology.

Parliamentary Corpus – structured sentence pairs from official proceedings.

Common Crawl Corpus – public web data, cleaned to remove foreign tokens, short or ungrammatical sentences.

Wikipedia Corpus – verified translated sentences from Wikipedia resources.

3.2 Model Choice

The selection of models was guided by open-source or open-weights availability to avoid licensing constraints and facilitate further development. Additionally, given memory constraints, around 2 billion parameter models were selected.

Gemma – Google's lightweight Gemini-based LLM for broad NLP tasks (Team et al., 2024).

EuroLLM – Unbabel's LLM, optimised for multilingual translation tasks across European languages (Martins et al., 2024).

Salamandra – BSC-LT's LLM, focused on European languages (Gonzalez-Agirre et al, 2025).

NLLB – Meta's NMT model, covering 200 low-resourced languages (Costa-Jussà et al., 2022).

Helsinki – NMT model specialised in Baltic languages, ideal for Lithuanian translation (Tiedemann et al., 2024).

Madlad – Google's NMT model supporting 400 languages (Kudugunta et al., 2023).

3.3 Evaluation Metrics

Model performance was quantitatively evaluated at two stages: baseline (pre-trained) and post-SFT. The following automatic metrics were used:

SacreBLEU – an improved version of BLUE, measuring n-grams overlap but limited in semantic meaning and synonyms (Papineni et al., 2002).

CHRF – based on character-level n-gram overlaps, effective for morphologically rich languages and correlating with human judgement (Popović, 2015; Lee et al., 2023).

ROUGE – evaluates precision and quality by measuring unigrams, bigrams, and sequence overlap (Lin and Och, 2004).

METEOR – enhances BLEU by considering synonym matching, stemming, and recall, accounting for a better semantic alignment (Banerjee and Lavie, 2005).

3.4 Human Evaluation

Translations were manually assessed on accuracy, fluency, and appropriateness, following Freitag et al. (2021) guidelines and scored from 1 (very poor) to 5 (excellent). Due to the time constraints, only a subset of sentences was evaluated that covered scientific, official, and casual contexts, with an emphasis placed on semantic ambiguity and metaphorical language. The aim of human evaluation was to identify the strengths and weaknesses of each model in producing

grammatically correct and contextually relevant translations.

3.5 Model Evaluation

Each model was configured to correctly handle source and target languages. NMT models require explicit language identifiers, such as appending a prefix to the input sentence for Helsinki. While LLMs are more general-purpose and use a prompt-based format. For EuroLLM, source and target prefixes were needed, where Gemma used special tokens for the start and end of inputs and responses.

Models translated 100 unique test sequences from the Flores+ dataset. The Transformers library was used for tokenisation and inference. Generated translations were decoded and compared using BLEU, METEOR, CHRF and ROUGE.

To assess the impact of model size, both ~2B and larger models (up to 9B parameters) were compared, excluding NLLB and Helsinki, as larger versions were not available. Apart from applying quantisation (4-bit) for efficiency, the evaluation process remained consistent with the previous step. Aiming to determine whether increasing model size shows improvement in translation quality.

3.6 Statistical and Practical Significance

To verify whether the differences in model performances were meaningful, t-scores were calculated for each metric using the formula:

$t - score = \frac{value - mean}{standard\ deviation}$.

Given a small sample size (6) and targeting a 95% confidence level, a t-critical value of 2.571 was used. Scores exceeding this threshold were considered to have a statistically significant difference (Benjamin et al., 2018).

Furthermore, to complement statistical significance, Cohen's d effect was used to evaluate the practical significance based on the formula: $Cohen's\ d = \frac{mean1 - mean2}{standard\ deviation}$

Providing the magnitude of the differences. Effects of up to 0.5 are considered small to medium, while values >0.8 indicate a strong effect (Gignac and Szodorai, 2016). Together, these measures ensure a robust and comprehensive interpretation of model performance differences.

3.7 Fine-Tuning Models

Due to high resource demands, fine-tuning focused on ~2B parameter models, utilising memory-efficient techniques. Models were quantised to 4-bit and fine-tuned with LoRA, targeting attention and feed-forward layers to reduce overhead while preserving performance.

Training used small batch sizes, combined with gradient accumulation, 2e-4 learning rate with linear scheduling, and Adam optimiser to produce gradual and efficient convergence. Models were evaluated consistently every 500 or 1,000 training steps with 100 test-set sentences separated from prior training and utilising the same metrics as in the baseline evaluation phase. This iterative process ensured steady performance monitoring and allowed parameter adjustments as needed.

4 Evaluation Results, Discussion and Error Analysis

4.1 Performance Comparison with Automatic Metrics

The pre-trained NMT models (Madlad, NLLB, Helsinki) generally outperformed LLMs (EuroLLM, Salamandra, Gemma).

Madlad presented the best BLEU, CHRF and overall scores, indicating strong alignment with reference translations. NLLB followed closely, maintaining a good balance between lexical accuracy and semantic variation (high METEOR). Helsinki performed well at the character-level (CHRF) despite a lower BLEU score. EuroLLM led amongst LLMs with relatively higher BLEU and METEOR scores. Gemma achieved the lowest overall scores with poor BLEU and ROUGE results, suggesting minimal overlap and improper sentence structure.

These findings point to NMT models being better suited for machine translation than LLMs.

Models	BLEU	METEOR	CHRF	ROUGE
Madlad	28	0.55	60	0.55
NLLB	26	0.52	58	0.52
Helsinki	21	0.48	55	0.48
EuroLM	19	0.42	49	0.43
Salaman.	17	0.41	50	0.42

Table 1: Pre-trained model evaluation results.

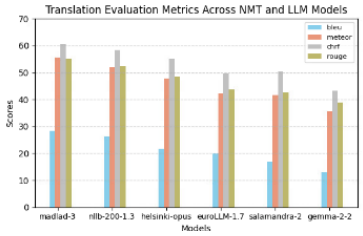

Figure 1: Pre-trained model evaluation.

4.2 Statistical and Practical Significance

T-scores showed that while Madlad consistently performed best and Gemma worst, neither deviated significantly from the mean, not exceeding the 2.571 threshold of 95% confidence. This indicates no statistical performance difference among models.

However, Cohen's d revealed strong practical differences contradicting the t-score. The effect sizes between the best- and worst-performing models, Madlad and Gemma, were on average 2.60 across all metrics, well above the 0.8 threshold for a large effect. Despite not reaching statistical significance, the practical performance difference was considerable.

4.3 Model Size Comparison

Larger models (~9B parameters) demonstrated consistent performance gain over their smaller variants (~2B), raising BLUE scores by 3-5 points while METEOR gains showed more variability, ranging from 0.03 to 0.12 points. The performance increase was more noticeable in LLMs, where NMT models benefited less from scaling up.

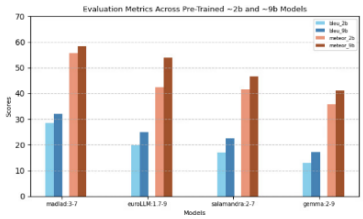

Figure 2: Pre-trained and Scaled Comparison.

Model	bleu_2	bleu_9	meteor_2	meteor_9
Madlad	28	31	0.55	0.58
EuroLLM	19	24	0.42	0.54
Salaman	17	22	0.41	0.46
Gemma	13	17	0.35	0.41

Table 2: Measures of Pre-trained and Scaled Models.

4.4 Pre-trained and Fine-tuned Comparison

Supervised fine-tuning demonstrated clear improvements across all models. BLEU scores rose by 5-8 points, with Madlad gaining 5 and Gemma 8. METEOR improved by 0.04-0.13, with the largest gains observed in LLMs (EuroLLM +0.10, Gemma +0.13).

While NMT models led with strong baseline performance, they showed moderate improvement. In contrast, LLMs started with lower scores but presented comparably larger gains, narrowing the performance gap. Overall, fine-tuning had the strongest impact on LLMs, significantly enhancing their translation quality.

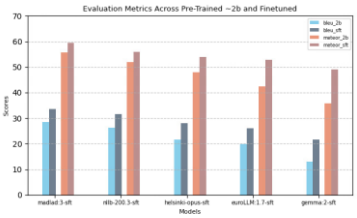

Figure 3: Pre-trained and Fine-tuned Comparison

Models	BLEU	METEOR	CHRF	ROUGE
Madlad	28	0.55	60	0.55
Madlad-lt	33	0.59	62	0.58
NLLB	26	0.52	58	0.52
NLLB-lt	31	0.55	60	0.55
Helsinki	21	0.48	55	0.49
Helsinki-lt	28	0.53	58	0.55
EuroLM	19	0.42	49	0.43
EuroLM-lt	26	0.52	62	0.55
Gemma	13	0.35	43	0.39
Gemma-lt	21	0.48	59	0.53

Table 3: Measures of Pre-trained and Fine-tuned.

4.5 Comparison of Fine-tuning and Scaling

Fine-tuning small models (~2B) led to substantial gains (BLEU +5-8, METEOR +0.04-0.13). However, compared to larger pre-trained models (~9B), fine-tuned models saw smaller performance differences (BLEU +1-4, METEOR +0.02-0.07).

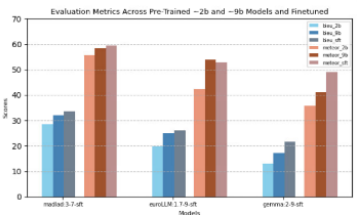

Figure 4: Scaled and Fine-tuning Comparison.

Model	bleu_9	bleu_ft	meteor_9	meteor_ft
Madlad	31	33	0.58	0.59
EuroLLM	24	26	0.54	0.52
Gemma	17	21	0.41	0.48

Table 4: Measures of Scaling and Fine-tuning.

4.6 Discussion on Pre-trained Model Evaluation

Automatic evaluation demonstrated that NMT models consistently outperformed LLMs in translation tasks, highlighting their domain-specific optimisation and better semantic handling. Madlad and NLLB taking the lead across all models, which could be attributed to their extensive multilingual capabilities, enabling broader linguistic variation and generalisation through parameter sharing (Pires et al., 2019). While the underlying success factor is the emphasis on data quality, where Madlad's team prioritised manual auditing, while NLLB curated custom corpus (Kudugunta et al., 2023; Costa-Jussà et al., 2022). Remarkably, NLLB achieved nearly identical results to Madlad despite having half the parameters, likely due to the use of back-translation and knowledge distillation techniques. Helsinki, despite being smaller (<1B), surpassed all LLMs, benefiting from its specialisation in Baltic languages. Nevertheless, it still trailed behind Madlad and NLLB, possibly because it was developed with limited resources compared to other company-backed models. EuroLLM and Salamandra performed comparably to Helsinki, showing that smaller LLMs can achieve competitive performance when designed with a task-specific focus and an emphasis on a high-quality, diverse dataset. Finally, Gemma produced the weakest results, despite its equivalent size and utilisation of distillation, likely caused by its English-focused training, which lacked multilingual depth (Team et al., 2024).

4.7 Discussion on Statistical and Practical Significance

The lack of statistical significance in the t-score could be attributed to the small sample size, which increased the probability of type II error (Huang, 2017). In comparison, Cohen's d practical value revealed a large effect (2.5) between best and worst models and a small effect between close performers (e.g., EuroLLM – Salamandra at 0.12), aligning with automatic evaluations. Though threshold values are estimated and may not be universal (Corell et al., 2020), when used alongside other evaluation methods, they reinforce the reliability of the study's findings.

4.8 Error Analysis with Human Evaluation

Models varied in accuracy with most common mistakes including literal translations of idioms e.g. "field" was translated as "physical location" instead of "area of research or "shine a light" lost its metaphorical meaning. Mistranslations of uncommon terms such as "rabid dog" was interpreted by Gemma as "red dog", while Helsinki presented "rabid" as "rabin". Hallucinations were observed from LLMs, particularly from EuroLLM, which regularly appended incorrect dates.

Fluency issues were widespread, with grammar being the most common error, with models using incorrect suffixes, verb tenses and pronouns. Notably, Gemma occasionally repeated words or used basic synonyms, showing limited vocabulary.

Appropriateness, which considers contextual and cultural relevance, proved that most models lacked official, scientific or field-specific terminologies and often reused English phrases.

Madlad: Strong domain-specific terms, though making minor grammatical errors. NLLB: Promising results but prone to ungrammatical and inaccurate terminology. Helsinki: Performed poorly with often mistranslations and Anglicisms. EuroLLM: Preserved the intent but suffered from hallucinations (e.g. added dates). Salamandra: Better than Gemma but had a limited vocabulary and common mistranslations. Gemma: Weakest among all models with frequent grammatical and terminology errors or untranslated phrases.

Models	Accur.	Flue.	Appr.	Total
Madlad	4	4	5	**4**
NLLB	3	4	4	**4**
EuroLLM	3	4	4	4
Helsinki	3	3	3	**3**
Salaman.	3	3	3	**3**
Gemma	2	2	2	**2**

Figure 5: Measures of Human Evaluation.

4.9 Discussion on Error Analysis

Human evaluation largely reinforced the automatic metrics rankings, while identifying overlooked word-level mismatches and error patterns. Where Madlad and NLLB correlated to automatic

evaluation, while Helsinki and EuroLLM diverged. EuroLLM exhibited frequent hallucinations, whereas Helsinki was more accurate but struggled with domain-specific terms. Gemma performed the worst with weak grammatical comprehension and word repetitions. These insights highlight the need for improved grammatical accuracy, contextual awareness, and vocabulary breadth in Lithuanian.

4.10 Overall Discussion of Results

Performance collectively improved with model size, while NMT models, due to specialised design and insufficient datasets, face a plateau in scaling effect (Kaplan et al., 2020; Ghorbani et al., 2021). In contrast, LLMs showed distinct improvement, accentuating better generalisation with increased parameters (Wei et al., 2024). However, scaling is a resource-intensive choice, making it impractical for low-funded research (Whittaker, 2021).

Fine-tuning offers a cost-effective alternative, significantly increasing smaller models' performance, especially LLMs, by enhancing the output's structure and coherence. Regardless of these benefits, this process has undesirable drawbacks, like concept forgetting, dependency on data quality and overfitting after a certain point (Mukhoti et al., 2023; Dodge et al., 2020). These issues are especially concerning in low-resource languages with limited diversity and quality data.

Moreover, smaller models (<2B) often lack sufficient multilingual representations (Conneau et al., 2020). While parameter-efficient methods such as LoRA can help, they cannot fully compensate for the advantages offered by large-scale models (Pfeiffer et al., 2020). Therefore, fine-tuning improves performance but does not overcome the inherent limitations of smaller models.

5 Conclusion

5.1 Research Limitations

This study was constrained by 12GB VRAM GPU (UcrelHex, 2024), which restricted the ability to fine-tune or evaluate larger models. Access to more powerful hardware may have yielded different results. Additionally, the focus on open-source/open-weight models ensured transparency and accessibility but excluded closed-source alternatives, possibly limiting performance range.

5.2 Future Work

Future research should prioritise expanding resources for low-resource language communities, as emphasised by the NLLB project, which focused on dataset collection before model design. With the use of distillation, to ensure efficiency, however, this process is limited by its knowledge retention and alternatives such as the Mixture of Experts (MoE) framework show promise by activating only the relevant networks, supporting scalability without increasing computational costs (Koishekenov et al., 2022).

Furthermore, advocating for open-source models is essential in supporting ethical, inclusive and transparent NLP research, especially in underrepresented languages. However, many high-performing models remain closed-source, limiting accessibility and collaboration (Worth et al., 2024).

As model architectures evolve, a clearer classification standard is needed as inconsistencies between model labelling complicate comparisons. Less ambiguous categorisation would enhance transparency and rationalise future research.

5.3 Overall Conclusion

This research evaluated LLMs and NMT models' performance in translating into Lithuanian, a low-resourced language, and revealed consistent outperformance of small NMT models compared to similarly sized LLMs. However, after scaling models (~7-9B parameters), higher performance gains were observed with LLMs, suggesting their better generalisation abilities while NMT models remain more efficient for translation tasks within resourced-constrained settings. Additionally, fine-tuning significantly enhances translation quality, introducing trade-offs as potential knowledge loss.

Ultimately, the key barriers to expanding translation capabilities for underrepresented languages remain computational constraints and data availability. Addressing these challenges requires continued investment in multilingual datasets and efficient training methods for building inclusive and reliable translation systems.

References

Baack, S., 2024, June. A critical analysis of the largest source for generative ai training data: Common crawl. In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (pp. 2199-2208). https://dl.acm.org/doi/10.1145/3630106.3659033

Bahdanau, D., 2014. Neural machine translation by jointly learning to align and translate. https://arxiv.org/abs/1409.0473

Banerjee, S. and Lavie, A., 2005, June. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization (pp. 65-72). https://dl.acm.org/doi/10.5555/1626355.1626389

Benjamin, D.J., Berger, J.O., Johannesson, M., Nosek, B.A., Wagenmakers, E.J., Berk, R., Bollen, K.A., Brembs, B., Brown, L., Camerer, C. and Cesarini, D., 2018. Redefine statistical significance. Nature human behaviour, 2(1), pp.6-10. https://doi:10.1038/s41562-017-0189-z

Brown, P.F., Cocke, J., Della Pietra, S.A., Della Pietra, V.J., Jelinek, F., Lafferty, J., Mercer, R.L. and Roossin, P.S., 1990. A statistical approach to machine translation. Computational linguistics, 16(2), pp.79-85. https://aclanthology.org/J90-2002/

Chakravarthi, R., B., Arcan, M., and McCrae, J., P., 2019 Comparison of Different Orthographies for Machine Translation of Under-Resourced Dravidian Languages. In 2nd Conference on Language, Data and Knowledge (LDK 2019). Open Access Series in Informatics (OASIcs), Volume 70, pp. 6:1-6:14. https://doi.org/10.4230/OASIcs.LDK.2019.6

Church, K.W., Chen, Z. and Ma, Y. (2021) 'Emerging trends: A gentle introduction to fine-tuning', Natural Language Engineering, 27(6), pp. 763-778. https://doi:10.1017/S1351324921000322

Correll, J., Mellinger, C., McClelland, G.H. and Judd, C.M., 2020. Avoid Cohen's 'small','medium', and 'large'for power analysis. Trends in cognitive sciences, 24(3), pp.200-207. https://doi.org/10.1016/j.tics.2019.12.009

Costa-jussà, M. R., et al. (2022). No Language Left Behind: Scaling Machine Translation for Low-Resource Languages. Proceedings of the 2022 Annual Conference on Neural Information Processing Systems. https://arxiv.org/abs/2207.04672

Das, B.C., Amini, M.H. and Wu, Y., 2025. Security and privacy challenges of large language models: A survey. ACM Computing Surveys, 57(6), pp.1-39. https://arxiv.org/abs/2402.00888

Dettmers, T., Pagnoni, A., Holtzman, A. and Zettlemoyer, L., 2024. Qlora: Efficient finetuning of quantized llms. Advances in Neural Information Processing Systems, 36. https://arxiv.org/abs/2305.14314

Dong, D., Wu, H., He, W., Yu, D. and Wang, H., 2015, July. Multi-task learning for multiple language translation. https://aclanthology.org/P15-1166/

Dong, J., 2024. Natural Language Processing Pretraining Language Model for Computer Intelligent Recognition Technology. ACM Trans. Asian Low-Resour. Lang. Inf. Process. https://doi.org/10.1145/3605210

Dong, L., Yang, N., Wang, W., Wei, F., Liu, X., Wang, Y., Gao, J., Zhou, M. and Hon, H.W., 2019. Unified language model pre-training for natural language understanding and generation. Advances in neural information processing systems, 32. https://arxiv.org/abs/1905.03197

Ghorbani, B., Firat, O., Freitag, M., Bapna, A., Krikun, M., Garcia, X., Chelba, C. and Cherry, C., 2021. Scaling laws for neural machine translation. https://arxiv.org/abs/2109.07740

Gignac, G.E. and Szodorai, E.T., 2016. Effect size guidelines for individual differences researchers. Personality and individual differences, 102, pp.74-78. https://doi.org/10.1016/j.paid.2016.06.069

Gonzalez-Agirre, A., Pàmies, M., Llop, J., Baucells, I., Da Dalt, S., Tamayo, D., Saiz, J.J., Espuña, F., Prats, J., Aula-Blasco, J. and Mina, M., 2025. Salamandra Technical Report. https://arxiv.org/html/2502.08489v2

Hsieh, C.Y., Li, C.L., Yeh, C.K., Nakhost, H., Fujii, Y., Ratner, A., Krishna, R., Lee, C.Y. and Pfister, T., 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. https://arxiv.org/abs/2305.02301

Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L. and Chen, W., 2021. Lora: Low-rank adaptation of large language models. https://arxiv.org/abs/2106.09685

Huang, H., 2017. Uncertainty estimation with a small number of measurements, part I: new insights on the t-interval method and its limitations. Measurement Science and Technology, 29(1), p.015004. DOI:10.1088/1361-6501/aa96c7

Koehn, P., Khayrallah, H., Heafield, K. and Forcada, M.L., 2018, October. Findings of the WMT 2018 shared task on parallel corpus filtering. In EMNLP 2018 Third Conference on Machine Translation (WMT18) (pp. 726-739). https://aclanthology.org/W18-6453/

Koishekenov, Y., Berard, A. and Nikoulina, V., 2022. Memory-efficient nllb-200: Language-specific expert pruning of a massively multilingual machine translation model. *https://aclanthology.org/2023.acl-long.198/*

Kudugunta, S., Caswell, I., Zhang, B., Garcia, X., Xin, D., Kusupati, A., Stella, R., Bapna, A. and Firat, O., 2023. Madlad-400: A multilingual and document-level large audited dataset. Advances in Neural Information Processing Systems, 36, pp.67284-67296. https://arxiv.org/abs/2309.04662

Kudugunta, S., Huang, Y., Bapna, A., Krikun, M., Lepikhin, D., Luong, M.T. and Firat, O., 2021. Beyond distillation: Task-level mixture-of-experts for efficient inference. *https://arxiv.org/abs/2110.03742*

Lee, S., Lee, J., Moon, H., Park, C., Seo, J., Eo, S., Koo, S. and Lim, H., 2023. A survey on evaluation metrics for machine translation. Mathematics, 11(4), p.1006. https://doi.org/10.3390/math11041006

Lin, C.Y. and Och, F.J., 2004, June. Looking for a few good metrics: ROUGE and its evaluation. In Ntcir workshop. https://aclanthology.org/W04-1013.pdf

Lopez, A., 2008. Statistical machine translation. ACM Computing Surveys (CSUR), 40(3), pp.1-49. https://doi.org/10.1145/1380584.1380586

Martins, P.H., Fernandes, P., Alves, J., Guerreiro, N.M., Rei, R., Alves, D.M., Pombal, J., Farajian, A., Faysse, M., Klimaszewski, M. and Colombo, P., 2025. Eurollm: Multilingual language models for europe. Procedia Computer Science, 255, pp.53-62. https://arxiv.org/abs/2409.16235

Mukhoti, J., Gal, Y., Torr, P.H. and Dokania, P.K., 2023. Fine-tuning can cripple your foundation model; preserving features may be the solution. *https://arxiv.org/abs/2308.13320*

Nakvosas, A., Daniušis, P., Mulevicius, V., 2024. Open Llama2 Model For The Lithuanian Language. A Preprint. Neurotechnology. https://arxiv.org/html/2408.12963v1

NLLB Team (2022) No Language Left Behind: Scaling Human-Centered Machine Translation. arXiv preprint arXiv:2207.04672. Available at: https://arxiv.org/abs/2207.04672

Pakray, P., Gelbukh, A. and Bandyopadhyay, S. (2025) 'Preface: Special issue on Natural Language Processing applications for low-resource languages', Natural Language Processing, 31(2), pp. 181–182. doi:10.1017/nlp.2024.34.

Papineni, K., Roukos, S., Ward, T., and Zhu, W., Z., 2002. BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL '02). Association for Computational Linguistics, USA, 311–318. https://doi.org/10.3115/1073083.1073135

Pires, T., Schlinger, E. and Garrette, D., 2019. How multilingual is multilingual BERT?. https://aclanthology.org/P19-1493/

Popović, M., 2015, September. chrF: character n-gram F-score for automatic MT evaluation. https://aclanthology.org/W15-3049/

Poupard, D., 2024. Attention is all low-resource languages need. Translation Studies, 17(2), pp. 424–427. https://doi.org/10.1080/14781700.2024.2336000

Sennrich, R., Haddow, B. and Birch, A., 2015. Improving neural machine translation models with monolingual data. https://aclanthology.org/P16-1009/

Sutskever, I., 2014. Sequence to Sequence Learning with Neural Networks. https://arxiv.org/abs/1409.3215

Team, G., Riviere, M., Pathak, S., Sessa, P.G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A. and Ferret, J., 2024. Gemma 2: Improving open language models at a practical size. https://arxiv.org/abs/2408.00118

Tiedemann, J., Aulamo, M., Bakshandaeva, D., Boggia, M., Grönroos, S.A., Nieminen, T., Raganato, A., Scherrer, Y., Vázquez, R. and Virpioja, S., 2024. Democratizing neural machine translation with OPUS-MT. https://arxiv.org/abs/2212.01936

Toral, A., Esplá-Gomis, M., Klubička, F. et al. Crawl and crowd to bring machine translation to under-resourced languages. Lang Resources & Evaluation 51, 1019–1051 (2017). https://doi.org/10.1007/s10579-016-9363-6

Vaswani, A., 2017. Attention is all you need. Advances in Neural Information Processing Systems. https://arxiv.org/abs/1706.03762

Vidler, J. and Rayson, P. (2024) *UCREL - Hex: A shared, hybrid multiprocessor system*. https://github.com/UCREL/hex

Wang, H., Wu, H., He, Z., Huang, L. and Church, K.W., 2022. Progress in machine translation. Engineering, 18, pp.143-153.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, QV., Zhou, D., 2022. Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems. https://arxiv.org/abs/2201.11903

White, M., Haddad, I., Osborne, C., Yanglet, X.Y.L., Abdelmonsef, A. and Varghese, S., 2024. The model openness framework: Promoting completeness and openness for reproducibility, transparency, and usability in artificial intelligence. https://arxiv.org/abs/2403.13784

Whittaker, M., 2021. The steep cost of capture. Interactions, 28(6), pp.50-55.

Worth, S., Snaith, B., Das, A., Thuermer, G. and Simperl, E., 2024. AI data transparency: an exploration through the lens of AI incidents. *https://arxiv.org/abs/2409.03307*

Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z. and Du, Y., 2023. A survey of large language models. https://arxiv.org/abs/2303.18223