Machine Translation in the AI Era: Comparing previous methods of machine translation with large language models

William Jock Boyd
School of Computing
and Communications
Lancaster University
United Kingdom
williamboyd106@gmail.com

Ruslan Mitkov
School of Computing
and Communications
Lancaster University
United Kingdom
r.mitkov@lancaster.ac.uk

Abstract

The aim of this paper is to compare the efficacy of multiple different methods of machine translation in the French-English language pair. There is a particular focus on Large Language Models given they are an emerging technology that could have a profound effect on the field of machine translation. This study used the European Parliament's parallel French-English corpus, testing each method on the same section of data, with multiple different Neural Translation, Large Language Model and Rule-Based solutions being used. The translations were then evaluated using BLEU and METEOR scores to gain an accurate understanding of both precision and semantic accuracy of translation. Statistical analysis was then performed to ensure the results validity and statistical significance. This study found that Neural Translation was the best translation technology overall, with Large Language Models coming second and Rule-Based translation coming last by a significant margin. It was also discovered that within Large Language Model implementations that specifically trained translation capabilities outperformed emergent translation capabilities.

1 Introduction

This study aims to compare previous and current methods of Machine Translation (MT) with Large Language Models (LLMs) to gauge the effectiveness of novel technologies in the field of MT.

The continuous improvement of technology in the MT space often leads to older methods being left behind – especially in the modern day as more and more companies make the pivot to LLMs. These previous methods, such as rules-based MT, can be effective in situations where there is a lack of resources available to train models. Additionally, LLMs trade off of accuracy for natural sounding translations could cause myriad issues in areas where accuracy is paramount such as the medical or legal fields. This suggests that Neural Machine Translation (NMT) is a better solution for these contexts at the moment. The question of LLMs in the field on MT is still in the early stages of being researched but does have promising results in early studies (Mujadia et al., 2023). However, these studies are often made with comparisons to other LLMs, giving no context as to how they perform against other methods of translation. Given the intense amount of resources required to train and run LLMs, an accurate inter-method comparison would allow potential users of these systems to evaluate the pros and cons before committing the time and resources required to use them.

This paper aims to create a clear picture on how rules-based translation, neural translation, and large language models and compare to each other on translation of the same text, and how different implementations of these methods can affect translation quality. This will give future research a baseline to compare from when progressing the field. This research is novel in that no other study has ever compared these three methods of MT in the same framework before. These new contributions will provide a clear picture of the current MT landscape giving insight as to where research should go in future. They will also let developers planning to use MT as part of their product to make an informed decision on which method is best for them, based on the trade-offs of each one. Within this study the efficacy of different translation approaches for LLMs will also be investigated, allowing developers of this technology to tailor their efforts depending on the task.

This study is highly relevant to the automatic extraction of socio-political events from text, given its focus on automatic translation and multilinguality. In multilingual contexts, translation methods are often essential for enabling such extraction. The data used in this research are drawn from the European Parliament's French–English parallel corpus, which provides extensive coverage of socio-

Proceedings of the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Texts associated with RANLP 2025, pages 38–51, Varna, Bulgaria, Sep 13, 2025.
https://doi.org/10.26615/978-954-452-099-1-005

political events. The findings of this research offer valuable guidance to researchers in selecting suitable approaches for tackling multilingual tasks of this kind.

2 Related Work

Although significant research on each of these methods has been done individually. And research on comparisons between systems like RBMT, NMT and LLMs has been done, these systems have never all been compared together using the same corpus with the same preprocessing on the translation results. Additionally, a majority of research only compares two types of systems at a time, whereas this study compares 3 types of systems with different implementations of those types. The rest of this subsection will discuss prior studies done on this topic, the limitations of that research and the significance of the research being done in this study.

Historical studies naturally covered RBMT and Statistical Machine Translation (SMT) systems such as this study by Costa-Jussà et al. (2012) comparing RBMT and SMT on Catalan – Spanish MT systems across 2 domains. This research was key in defining performance differences between systems. Another key paper evaluating direct performance comparisons between the two systems is the paper by S and Bhattacharyya (2017) which uses the Marathi–Hindi language pair. This is a study with very different takeaways due to the structural differences between Marathi and Hindi, compared to the very similar languages of Catalan and Spanish. In more modern research NMT models have started to be included as part of these studies with multiple studies being published in comparing all 3 system types by De Silva and Hansadi (2024) and Dwivedi et al. (2025) covering this area of research. Additionally, as LLMs have started displaying more and more translation capabilities comparison with existing NMT solutions has started to be done. Such as a paper by Sizov et al. (2024) comparing NMT, LLM, and human translations using human and automatic evaluation. This study sets itself apart by comparing the translations LLMs produce to other systems outputs, rather than focusing on the technique specifically used to get the LLMs to produce this output. However, all of these studies have limitations which will be addressed in the next section.

In studies done less recently only two different methods were compared, not allowing a complete and fair comparison across multiple different systems. This aspect did change with the advent of NMT as researchers wanted to see how it would match up with pre-existing techniques. After the introduction of LLMs to the MT space this focus has narrowed again as studies look to see how LLMs match up against the latest and greatest technologies on offer, rather than how they fit amongst all the available technologies. Additionally according to a meta-analysis by Marie et al. (2021) BLEU scores have been used ineffectively. As studies often copied scores directly from other research without any consideration for how the score was calculated, rendering the comparison invalid. Additionally, without statistical significance testing, the difference between the two scores could be completely coincidental, this is an important tool that is rarely used and the usage of which has been declining over time.

3 Methodology

This section will cover the design of the experiment carried out, including the questions to be answered by the experiment; the translation systems being evaluated and any configurations required to make them work; the corpus these translation systems were tested on; the evaluation metrics use; their specific implementations and the statistical analysis methods used. The experiment protocol will then be discussed, with an evaluation of how these protocols ensure fair comparison and an explanation of how the scores were calculated and compared.

3.1 Corpus Selection and Preparation

The corpus used was the European Parliament's French-English parallel corpus (Koehn, 2005). This was chosen as it covers a variety of domains with discussions ranging from law to the medical field, to nature conservation. This variety enables an excellent insight into how MT systems perform across multiple domains. In addition, the size of this corpus allows for ample development and experiment sets, meaning the development of the testing systems can emulate the experiment itself more closely in terms of scale, without restricting the size of the experiment data. The first 40,000 lines of the last 10% of the data were used as a development set to ensure the integrity of the data, then the next 60,000 lines made up the experiment data set. The only preprocessing done on the data set was to remove unreasonably long sentences that would

exceed the token limits of the models being used.

3.2 Rule-based Model

Given the lack of freely available rule-based models, the only model evaluated in this study is Apertium (Forcada et al., 2011), an open-source RBMT toolkit. For this study version 2.9.4 of the base Apertium model, the English Apertium version, and the French Apertium version were installed. Then the French English language data from the Github was downloaded and the instructions there were used to install and set up the pair. To access the system the command line was used running a shell script that would split the complete experiment file into chunks. Apertium would process each chunk then the translations would be recombined in order. Apertium was chosen in this study as it is the most accessible RBMT model and has been used in multiple research studies previously (Costa-Jussà et al., 2012); (Corbí-Bellot et al.).

3.3 Neural models

Three neural models were assessed in this study to allow different styles and implementations to be evaluated against LLMs, enabling a better overall picture of how they fit in the space. All models were run locally with Huggingface's transformers library (Wolf et al., 2020) using the pipeline interface in Python to send data to the models and receive outputs. The largest models possible were used, as generally the larger the model, the better it performs. Every model was used in the default configuration, with the source languages being specified as French and the target language as English. The neural models chosen as part of this study are the following:

The Marian NMT system is a purely NMT system that uses the transformer architecture (Junczys-Dowmunt et al., 2018), it was developed as an efficient C++-only implementation of the architecture detailed in the paper "Attention is All You Need" (Vaswani et al., 2017). The particular version used was the French-English model from Opus MT (Tiedemann and Thottingal, 2020); (Tiedemann et al., 2022), which is a Marian model trained on the Opus parallel corpus.

Meta's M2M100 model (Fan et al., 2020) is a multilingual translation model that supports translation across 100 different languages. It still uses the same attention mechanism proposed by Vaswani et al. but only requires one model to translate between all these languages. M2M100 was created to address the traditional "English-Centric" approach of multilingual translators, which typically involves translating the source language into English, then English into the target language. The version used in this study was the 1.2 billion parameter version in order to enhance accuracy.

Meta's No Language Left Behind (NLLB) model (Team et al.) is another multilingual translation model but it supports many more languages. NLLB supports 200 different languages, with 150 of them being low-resource languages. The specific model used in this study is the 1.3 billion parameter version, the goal was to use the 3.3 billion parameter version but due to computing resource constraints, this option could not be used.

3.4 Large Language Models

Two LLMs were evaluated in this study to assess how different approaches towards translation capabilities in LLMs can change their effectiveness. Both models were run locally using Huggingface's transformers library and pipeline interface. The LLMs chosen for evaluation are the following:

Google's Text-to-Text Transfer Transformer or T5 (Raffel et al., 2023) is a large language model that treats every NLP task as a text-to-text problem[1]. This approach means T5 can in effect switch modes; this allows the system to approach translation as a task it was directly trained for, rather than as an emergent capability. The uniqueness of T5's approach positions it in a middle ground between NMT systems that can only translate and LLMs that are not trained for translation whatsoever. This technique significantly improves T5's ability to follow instructions and perform zero-shot tasks, allowing T5 to perform in this study despite the constrained computing resources. The specific model version was the FLAN-T5 large, an instruction-tuned version of T5. The model was used in its default configuration with the maximum number of new tokens it was allowed to produce set to 256. When translating, the model was prompted with "Translate from French to English" followed by the sentence to be translated.

Meta's Large Language Model Meta AI (Llama) (Touvron et al., 2023) is an open-source family of LLMs that aims to democratise AI access and enable research advancement. They are a more

[1]Many researchers consider T5 a Deep Learning model not an LLM. For the purposes of this study, T5 will be classed as an LLM due to its generation capabilities. Additionally, the use of T5 large gives more LLM like behaviour.

standard style of LLM being decoder only and pre-trained on large text corpora, meaning translation is an emergent capability. The Llama version used in this study was Llama 3.2 instruct with 3 billion parameters (Grattafiori et al., 2024). The instruct version is fine-tuned on instruction following data, this will improve the model's adherence to the translation request but not the translation itself. The configuration of the model was set to a maximum of 300 new tokens the precision of the model had to be reduced from 32-bit to 16-bit due to resource constraints. This causes a small reduction in overall accuracy, particularly in more nuanced expressions, but is necessary given the constraints of the experiment. The model was also set to only return the response to the prompt. To prompt the model, lists of dictionaries with role and content sections were used. The prompt used was "You are a French to English Translator, translate the input sentences and only give the output sentence" in the system role to set up the model, then in the user role the sentence was given to the model to be translated.

3.5 Evaluation Metrics

Two automated evaluation metrics were used in this study, BLEU score (Papineni et al., 2002) and METEOR score (Banerjee and Lavie, 2005). This approach was used as BLEU score alone can lead to incorrect conclusions about which systems are better according to a meta-evaluation of MT research (Marie et al., 2021), using METEOR avoids this pitfall and also evaluates the systems from a semantic perspective. The Python Natural Language ToolKit (NLTK) (Bird et al., 2009) implementations of both these scores were used. The reference translations for systems to be evaluated against were taken from the Europarl parallel corpus and no modifications were made to the reference translations.

To calculate BLEU score for each translation, the score for each sentence was calculated using the sentence_bleu() function in NLTK, then all the scores were averaged to get an overall score for the translation. Each n-gram was weighted equally, and no smoothing function was used. Sentence level BLEU calculation was used so that bootstrapping could be performed as part of the statistical analysis.

To calculate the METEOR score for each translation, the score for each sentence was calculated using the single_meteor_score() function as there was only one hypothesis per reference translation. The default parameter settings for this implementation were used as they have been studied and calibrated to align with human judgements.

3.6 Data

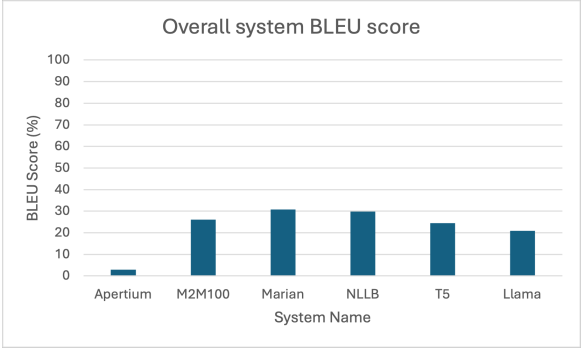

Figure 1: Graph of overall BLEU score for each system

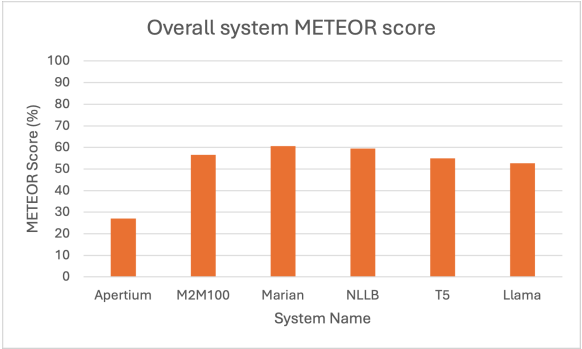

Figure 2: Graph of overall METEOR score for each system

System name	BLEU Score	METEOR Score
Apertium	2.86	27.07
M2M100	26.03	56.49
Marian	30.8	60.59
NLLB	29.8	59.44
T5	24.47	54.97
Llama	20.88	52.6

Figure 3: Table of overall BLEU and METEOR scores for each system

3.7 Statistical Analysis

In order for any conclusions to be made about the results their statistical significance needs to be evaluated to demonstrate they are not just random chance. A meta-evaluation Marie et al. (2021) found that only a minority of papers perform statistical significance testing on their scores. This study addresses this methodological gap by using

bootstrap resampling to ensure the significance of the scores.

Bootstrap resampling was used to create 1000 samples for each system by randomly selecting the sentence level scores from each system with replacement. The size of each sample was 60000 scores - the same size as the original dataset for each system. The overall average of each sample was then recorded so that distributions of these averages could be made and evaluated for each system.

From these distributions, the mean, standard deviation, 95% confidence intervals, minimum, and maximum were calculated for each system. The mean is the primary performance metric and the number that will be compared between systems. The standard deviation shows the variability across samples and how consistent each system's performance is. The 95% confidence intervals establish a range where the true score for each system likely exists. The minimum and maximum values show the best and worst bootstrap samples - a small gap between these two numbers indicates less variability. Together, these metrics give a comprehensive statistical profile of each system's performance. Excel's standard functions were used to calculate these metrics;

The AVERAGE() function was used to calculate the mean of the bootstraps for each system.

STDEV.S() was used to calculate the standard deviation.

The 95% confidence intervals were calculated using the CONFIDENCE.NORM() function, which determines the margin of error based on standard deviation, alpha level, and sample size. An alpha level of 0.05 indicates a 95% confidence interval.

The MIN() and MAX() functions were used to calculate the minimum and maximum sample values for each system.

To compare each system with every other system meaningfully, the p-values between each system were calculated. To calculate the p-values the official result between two systems is compared with every bootstrap sample to see if they match, i.e. if in the official results one system is better; is it better in each bootstrap sample? If less than 5% of the bootstrap results contradict the original finding, meaning $p < 0.05$, then the official result is statistically significant. Calculation of these values was done in excel. To calculate the p-values between systems a formula was implemented to count the number of occurrences where the bootstrap result matched the actual comparison result between the two systems. A sum of these occurrences was then done, that sum was divided by 1000 and taken away from 1 to get the final value. The formula is as follows:

1 - SUM(IF(system 1 bootstrap values > system 2 bootstrap values, 1, 0))/1000

3.8 Apertium Translation

As Apertium cannot handle a huge number of lines at once, the translation data was split into chunks of 150 lines with each chunk in its own file. Apertium was then given each line from each file to translate, every 50 sentences the translations would be written to a file, giving 3 new translated files for every chunk. This was done for every file, and then the files were recombined to create a sentence-aligned file containing every translated sentence.

3.9 Neural Translation

The neural translation was all done from within one file, with each translator translating the source file sequentially so each could have the maximum compute resources available to it. Each system was set up using Huggingface's pipeline in translation mode, and then a for loop iterating through each line in the file was started, yielding each line to the pipeline, the corresponding output was then written to the results file in the same order as the source file ensuring sentence alignment.

3.10 LLM Translation

The LLM translation was done from two separate files given their need for slightly different setups and prompt structuring. T5 was also implemented with Huggingface's pipeline module, it was set up in text-to-text generate mode, with the number of max new tokens allowed to be generated each time set to 256 due to memory constraints. The same for loop to iterate through each line in the source file was used, the input to the model was "Translate from French to English *sentence to translate*". For Llama's implementation, an identical process was used, however, Llama was set up with a max new token count of 300 and 16-bit precision.

3.11 Score Calculation

To calculate the overall METEOR and BLEU scores for each system the individual score of both types for each sentence was calculated then an average of all these sentences was calculated to get

the overall score for each system. To perform bootstrapping a score was randomly selected from the original population of the 60,000 sentence-level scores and added to a new sample but left in the original population. The overall average for the sample was then calculated and added to a list of bootstrap averages for that system.

3.12 Summary

The comprehensive, robust approach detailed in this chapter shows that this experiment is competently able to answer the research questions posed. With a strong framework designed to effectively evaluate each method against the other, using multiple implementations of methods to gain a comprehensive understanding of the performance of each. The use of the Europarl corpus provides a diverse and well-established dataset for translation tasks. The use of both BLEU and METEOR provides a more thorough analysis of the translation quality of each system, as one evaluates the accuracy of the sentence and the other evaluates the semantic meaning. Additionally, statistical analysis with bootstrapping validates the significance of these results - ensuring that conclusions drawn from this study are reliable.

4 Discussion

This section will analyse the results presented in the previous section and discuss their implications.

4.1 Primary Research Question 1

How do RBMT, NMT, and LLM translation approaches compare across general translation tasks in the French-English language pair?

The initial comparison is quite clear with NMT coming out on top with the highest performing NMT system, Marian, having a BLEU score of 30.8% and a METEOR score of 60.59% (3), NMT was then followed by LLMs with the highest scoring LLM, T5, having a BLEU score of 24.47% and a METEOR score of 54.97% (3). RBMT was then last with a significantly larger gap as Apertium had a BLEU score of 2.86% and a METEOR score of 27.07% (3). This huge gap of nearly 28% in BLEU and nearly 33% shows the significant advancements that have been made in the space since the creation of Apertium. In particular, the larger disparity in METEOR scores shows NMT's ability to maintain semantic coherence over the whole translation compared to RBMT. Given NMT's dominance in the study, a comparison between them provides insight into which implementation provides the best translation. The best system was Marian, followed by NLLB with a BLEU score of 29.8% and a METEOR score of 59.44% (3) with M2M100 coming last in the category with a BLEU score of 26.03% and a METEOR score of 56.49% (3). Both systems tuned to translate multiple languages rather than just one pair performed worse than the system only trained for the French-English language pair, showing that even though good results can be achieved with a generalised system, specially trained systems will outperform.

4.2 Primary Research Question 2

Are LLMs the method that will become the prevailing technology in the translation space in future?

The results of this experiment indicate that LLMs cannot quite attain the level of translation accuracy of NMT models – whether they are multilingual or single-language systems. With a small performance difference between the lowest performing NMT model M2M100 (BLEU: 26.03%, METEOR: 56.49%) (3) and the highest performing LLM T5 (BLEU: 24.47%, METEOR: 54.97%) (3) of around 1.5% across both scores. Despite these small differences, the comparison is significant due to the p-value of 0 (16,17) between these systems. When comparing between best-performing NMT system Marian (BLEU: 30.8%, METEOR: 60.59%) (3), and the worst-performing LLM Llama (BLEU: 20.88%, METEOR: 52.6%), there is BLEU gap of nearly 10% and a METEOR gap of nearly 8%. These score differences show that different implementations of LLMs using different approaches can drastically alter translation quality, paving the way for new LLM approaches to be used in the future. Consideration must also be made for LLMs' ability to perform general tasks beyond translation such as text generation, these extra facilities could lead to users taking a small hit in translation quality to have a single solution for all their problems, rather than dedicated systems for each task. However, LLMs incredibly high resource costs for similar or worse translation results limits their ability to spread as a translation tool, as training and running them requires huge time and infrastructure investments. In time, LLMs should become the prevailing technology as customers who already use LLMs will want translation capabilities included. NMT and LLM approaches may also be combined in a similar vein

to how T5 works, allowing for the translation quality of NMT systems, alongside the other abilities of LLMs.

4.3 Secondary Research Question

In the category of LLMs, how do the emergent capabilities of LLMs which have not been trained to do translation tasks compare to the capabilities of LLMs which have explicitly been trained to do translation tasks? Using T5 as the model explicitly trained for translation and Llama as the model with emergent capabilities it is clear there is a significant difference in translation quality between the two. T5's scores (BLEU: 24.47%, METEOR: 54.97%) are higher than Llama's (BLEU: 20.88%, METEOR: 52.6%) with the larger difference in BLEU score of nearly 5% compared to the difference in METEOR score of just over 2%. This gap between translation scores shows specialised training for an LLM significantly enhances translation precision while only slightly enhancing overall translation quality. This suggests that for situations where accuracy of translation is paramount, specifically trained LLMs are a better fit as they will better convey the meaning of the source text.

4.4 MT in Specialised Domains

These results can be extrapolated to gain insights into how these technologies would perform in different situations, such as in specific translation domains like legal or medical disciplines. In these domains translation precision and accuracy are paramount as errors can have serious consequences, as such the systems with the best scores overall, and particularly higher BLEU scores, would fare best in these domains. NMT systems, Marian in particular, are the solution for this given their top overall performance and high BLEU scores, indicating good precision. However, in domains that require less precision and more natural-sounding translations such as creative content like advertising, LLMs could play a key role. If creative companies are already using LLMs for other purposes, their ability to provide good translations that maintain semantic accuracy in an area where precision doesn't matter as much provides these companies with one technical solution for multiple areas.

4.5 Future Developments

As the MT technologies progress, it is important to distinguish which technologies will dominate in the near and long term. In the near term, NMT will remain the dominant technology as its significant performance advantage over other technologies suggests it will be the default choice for the highest-quality translation in the immediate future. In the long term, the best of both LLM and NMT technologies will likely converge, indicated by the small gap between LLM and NMT performance. This idea is also demonstrated by T5's approach of being trained for language translation on top of its general LLM capabilities, incorporating the strengths of both these technologies. As these technologies develop, the trade-off of functionality and computing cost will be prioritised over translation quality as it becomes less of a factor. The large computing costs but extra functionalities of LLMs need to be considered against NMT's lower computing costs but single functionality. Additionally, single-language pair NMT systems will start to be phased out as the close performance gap between Marian and NLLB of less than 1% indicates that multilingual NMT solutions will have equal performance to single-pair solutions.

4.6 Limitations of Analysis

To properly contextualise the analysis made in this section it is important to highlight the limitations of the study that produced the results. The use of automated evaluation metrics without any human evaluation can potentially cause false confidence as there is evidence to show that METEOR and BLEU can miss essential sentiment mistakes in translation (Saadany and Orasan, 2021). In addition, testing on a single high-resource language pair like French-English puts corpus-based translation systems at an advantage as the resources to train them properly, whereas RBMT systems often perform better with low-resource languages (Bayón and Sánchez-Gijón, 2019). The resource constraints in this project could have hindered LLM performance, particularly in the case of Llama, as the precision had to be reduced to 16-bit due to memory constraints and a model with fewer parameters was used. Despite these limitations, the statistical significance of the performance differences shows that the results discussed in this chapter are reliable. These constraints should be considered when interpreting the results of this study and applying its findings.

4.7 Summary

The key findings from this study answer the first research question, definitively showing that NMT is

the superior technology in both semantic accuracy and precision of translation. LLMs closely followed with much lower precision but were closer in semantic accuracy due to their ability to understand the structures of human language with RBMT coming last by a significant amount because of its inability to include semantic context when translating. Despite not being the top-performing technology the results shown by LLMs in this study were very promising, positioning them to become the prevailing technology in the MT field in future, especially when specially trained for translation tasks alongside generative capabilities. Within the LLM field, two different styles of translation were evaluated, emergent translation capabilities and LLMs trained for translation in the form of Llama and T5. T5 had better overall translation quality with a much bigger improvement over Llama in precision, showing that while emergent capabilities are impressive and could be used for non-critical translation, if accurate, precise translation is needed specially trained systems are better. These comparisons can be made with confidence due to the extensive statistical significance testing performed as part of this study, with every p-value being 0 the comparisons between each system are extremely statistically significant and can be evaluated as extremely valid. This study is the first to compare these three translation technologies and as a result, provides unique insight for users or developers considering implementing one of them.

5 Conclusion

The significance of this research is that there is a comprehensive evaluation framework comparing three different MT translation technologies to ensure the accuracy of results and comparison. These translations are also evaluated on both a word-by-word basis and overall semantic basis using multiple evaluation metrics, something many studies lack. The translation task itself covers multiple domains, allowing a true demonstration of each system's more diverse capabilities. The study also implements statistical analysis suggestions by Marie et al. in order to ensure the significance of the findings, leading to confidence that these results can be used to make informed decisions when using these systems in future. The development of a framework like this provides a consistent benchmark future technologies can be measured against. This paper also offers key insights into the current MT space and its potential future trajectory. The results of this study are directly relevant to the automatic extraction of socio-political events in multilingual contexts, where the use of automatic translation methods may be necessary.

5.1 Limitations

Despite this project's successes in creating effective results, multiple resource constraints limited the scope of the research. Computing restraints lead to smaller models being used – particularly when it came to LLMs – with Llama's 3.3 billion parameter model having to be used, despite the availability of larger models. Llama's precision also had to be reduced due to memory constraints with the hardware used. The use of the French-English language pair also favours data-driven approaches as it is a high-resource language pair with plenty of data available to train systems that need it. If this paper were to be repeated with more time allocated more language pairs from different language families would be added to assess the efficacy of each system with different grammatical structures and vocabularies. Statistical models would also be assessed to provide even more context of how different technologies perform.

5.2 Future Work

Future work directly stemming from this research could involve creating both broader and more specific studies. Future research projects with access to more compute or paid APIs can use larger, more performant models such as LLMs with 100 billion or more parameters. This allows better insight into very current technologies in a way that is unavailable with open-source resources. Another avenue of research developed from this would be repeating the same study with more RBMT systems on low-resource languages. This reverses the dynamic of corpus-based systems having an advantage allowing RBMT to show its use in more niche scenarios. A final branch of study resulting from this project would be developing and investigating hybrid NMT-LLM approaches to translation. These would also have to be evaluated from an LLM perspective to ensure the different training method would not affect its generative capabilities. This research would heavily advance the field of MT potentially removing the need for compromise.

Acknowledgements

This work has been partially supported by the CIDEXG/2023/12 project, funded by the Generalitat Valenciana

References

Satanjeev Banerjee and Alon Lavie. 2005. 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*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

María Do Campo Bayón and Pilar Sánchez-Gijón. 2019. Evaluating machine translation in a low-resource language combination: Spanish-Galician. In *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*, pages 30–35, Dublin, Ireland. European Association for Machine Translation.

Steven Bird, Ewan Klein, and Edward Loper. 2009. Natural language processing with Python: analyzing text with the natural language toolkit.

Antonio M Corbí-Bellot, Mikel L Forcada, Sergio Ortiz-Rojas, Juan Antonio Pérez, Gema Ramírez-Sánchez, Felipe Sánchez-Martínez, Iñaki Alegria, and Kepa Sarasola. An Open-Source Shallow-Transfer Machine Translation Engine for the Romance Languages of Spain.

Marta R. Costa-Jussà, Mireia Farrús, José B. Mariño, and José A. R. Fonollosa. 2012. Study and Comparison of Rule-Based and Statistical Catalan-Spanish Machine Translation Systems. *Computing and Informatics*, 31(2):245–270. Number: 2.

D. I. De Silva and D. G. P. Hansadi. 2024. Enhancing machine translation: Cross-approach evaluation and optimization of rbmt, smt, and nmt techniques. In *2024 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES)*, pages 1–8.

Ritesh Kumar Dwivedi, Parma Nand, and Om Pal. 2025. Hybrid NMT model and comparison with existing machine translation approaches. *Multidisciplinary Science Journal*, 7(4):2025146–2025146.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond English-Centric Multilingual Machine Translation. ArXiv:2010.11125 [cs].

Mikel L. Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M. Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seo-

hyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The Llama 3 Herd of Models. ArXiv:2407.21783 [cs].

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast Neural Machine Translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–

121, Melbourne, Australia. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Benjamin Marie, Atsushi Fujita, and Raphael Rubino. 2021. Scientific Credibility of Machine Translation Research: A Meta-Evaluation of 769 Papers. ArXiv:2106.15195 [cs].

Vandan Mujadia, Ashok Urlana, Yash Bhaskar, Penumalla Aditya Pavani, Kukkapalli Shravya, Parameswari Krishnamurthy, and Dipti Misra Sharma. 2023. Assessing Translation capabilities of Large Language Models involving English and Indian Languages. ArXiv:2311.09216 [cs].

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 on Association for Computational Linguistics*, ACL '02, pages 311–318, USA. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. ArXiv:1910.10683 [cs].

Sreelekha. S and Pushpak Bhattacharyya. 2017. Comparison of smt and rbmt; the requirement of hybridization for marathi-hindi mt.

Hadeel Saadany and Constantin Orasan. 2021. BLEU, METEOR, BERTScore: Evaluation of Metrics Performance in Assessing Critical Translation Errors in Sentiment-oriented Text. In *Proceedings of the Translation and Interpreting Technology Online Conference TRITON 2021*, pages 48–56. ArXiv:2109.14250 [cs].

Fedor Sizov, Cristina España-Bonet, Josef Van Genabith, Roy Xie, and Koel Dutta Chowdhury. 2024. Analysing translation artifacts: A comparative study of LLMs, NMTs, and human translations. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1183–1199, Miami, Florida, USA. 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, Jeff Wang, and Meta Ai. No Language Left Behind: Scaling Human-Centered Machine Translation.

Jörg Tiedemann, Mikko Aulamo, Daria Bakshandaeva, Michele Boggia, Stig-Arne Grönroos, Tommi Nieminen, Alessandro Raganato, Yves Scherrer, Raul Vazquez, and Sami Virpioja. 2022. Democratizing Neural Machine Translation with OPUS-MT. ArXiv:2212.01936 [cs].

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – Building open translation services for the World. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. ArXiv:2302.13971 [cs].

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. ArXiv:1706.03762 [cs].

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 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, Online. Association for Computational Linguistics.

A Bootstrap distributions and statistics tables

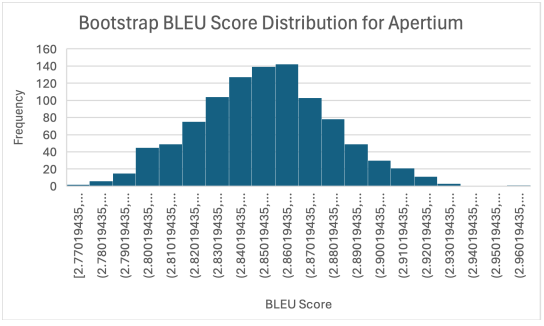

Figure 4: Apertium BLEU score bootstrap distribution

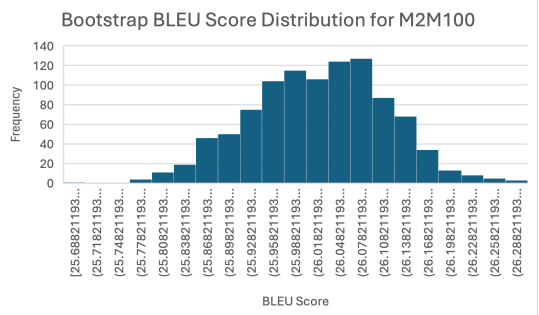

Figure 5: M2M100 BLEU score bootstrap distribution

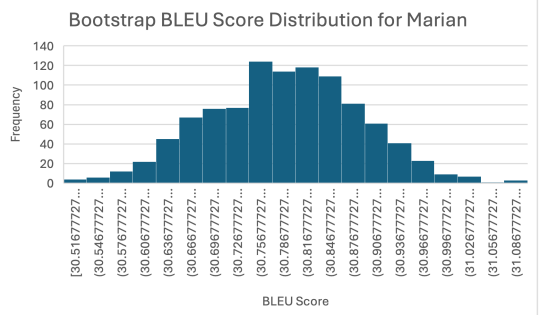

Figure 6: Marian BLEU score bootstrap distribution

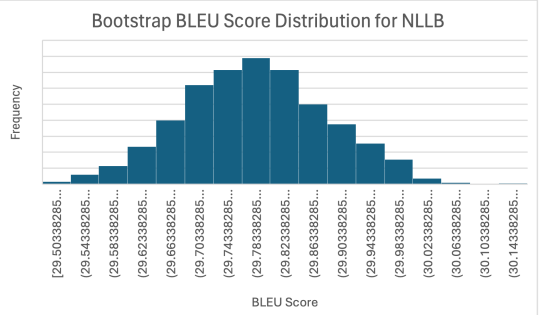

Figure 7: NLLB BLEU score bootstrap distribution

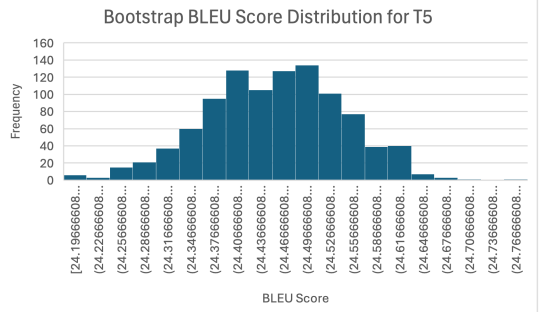

Figure 8: T5 BLEU score bootstrap distribution

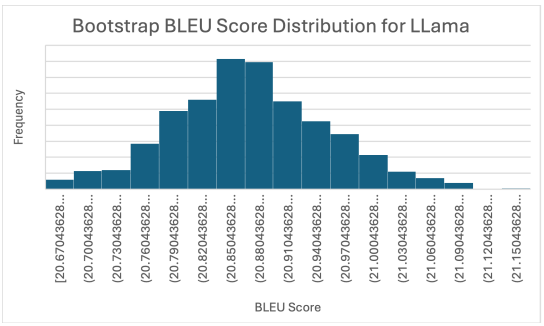

Figure 9: Llama BLEU score bootstrap distribution

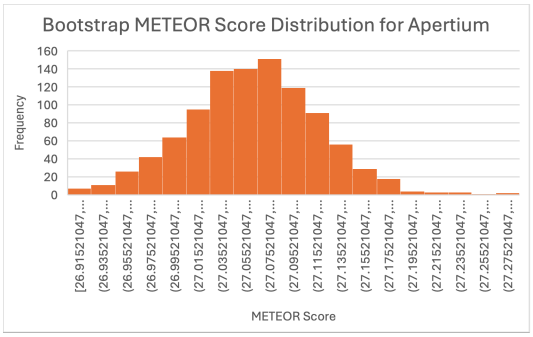

Figure 10: Apertium METEOR score bootstrap distribution

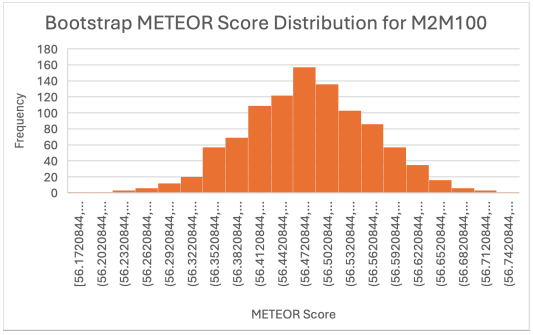

Figure 11: M2M100 METEOR score bootstrap distribution

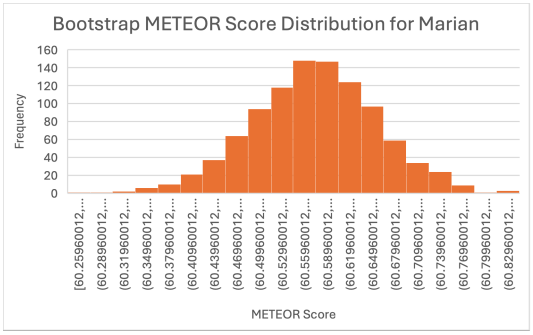

Figure 12: Marian METEOR score bootstrap distribution

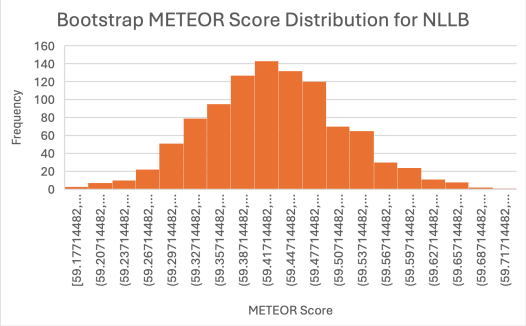

Figure 13: NLLB METEOR score bootstrap distribution

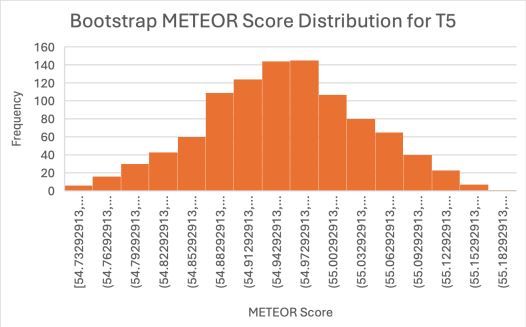

Figure 14: T5 METEOR score bootstrap distribution

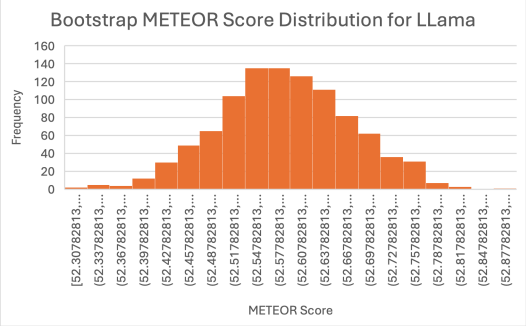

Figure 15: Llama METEOR score bootstrap distribution

Metrics	Llama bleu	M2M100 bleu	Marian bleu	NLLB bleu	T5 bleu	Apertium bleu
mean	20.8855	26.0353	30.8024	29.8023	24.4704	2.8556
std dev	0.0822	0.0919	0.0987	0.1013	0.0888	0.0290
CI	0.0051	0.0057	0.0061	0.0063	0.0055	0.0018
min	20.6704	25.6882	30.5168	29.5034	24.1967	2.7702
max	21.1543	26.2986	31.1079	30.1709	24.7679	2.9626

Figure 16: Statistics table for BLEU bootstrap scores

Metrics	Llama meteor	M2M100 meteor	Marian meteor	NLLB meteor	T5 meteor	Apertium meteor
mean	20.8853	52.6005	60.5876	59.4417	54.9659	27.0714
std dev	0.0823	0.0876	0.0842	0.0873	0.0837	0.0552
CI	0.0051	0.0054	0.0052	0.0054	0.0052	0.0034
min	20.6704	52.3078	60.2596	59.1771	54.7329	26.9152
max	21.1543	52.8834	60.8371	59.7371	55.1862	27.2814

Figure 17: Statistics table for METEOR bootstrap scores

B Link to project github

https://github.com/boydw27/MTInTheAIEra