Error Analysis of Multilingual Language Models in Machine Translation: A Case Study of English-Amharic Translation

Hizkiel Mitiku Alemyehu and **Hamada M. Zahera** and **Axel-Cyrille Ngonga Ngomo**
Department of Computer Science, Paderborn Universtiy
hizkiel.alemayehu@uni-paderborn.de hamada.zahera@uni-paderborn.de
axel.ngonga@uni-paderborn.de

Abstract

Multilingual large language models (mLLMs) have significantly advanced machine translation, yet challenges remain for low-resource languages like Amharic. This study evaluates the performance of state-of-the-art mLLMs, specifically NLLB-200 (NLLB3.3, NLLB1.3 Distilled1.3, NLB600) and M2M (M2M1.2B, M2M418), in English-Amharic bidirectional translation using the Lesan AI dataset. We employed both automatic and human evaluation methods to analyze translation errors. Automatic evaluation used BLEU, METEOR, chrF, and TER metrics, while human evaluation assessed translation quality at both word and sentence levels. Sentence-level accuracy was rated by annotators on a scale from 0 to 5, and word-level quality was evaluated using Multidimensional Quality Metrics. Our findings indicate that the NLLB3.3B model consistently outperformed other mLLMs across all evaluation methods. Common error included mistranslation, omission, untranslated segments, and additions, with mistranslation being particularly common. Punctuation and spelling errors were rare in our experiment.

1 Introduction

Machine Translation (MT), the automated process of converting text from one language to another, has shown significant advancements in recent years (Vaswani et al., 2017; Haddow et al., 2022). However, these improvements have primarily focused on high-resource languages. In contrast, low-resource languages, such as Amharic language –the official language of Ethiopia, spoken by 20 million people–, have not received the same level of attention (Tesfagergish et al., 2022).

Previous research on Amharic machine translation has primarily focused on fine-tuning language models. However, few studies have assessed the effectiveness of multilingal language models for low-resource languages such as Amharic. For example, Teshome et al. (2015) employed a statistical method to translate texts, improving a baseline phrase-based system. Similarly, Gezmu et al. (2021) addressed inflectional morphology by using subword units and adjusting model parameters to better suit low-data environments. Their approach effectively handled Amharics morphological complexity. Biadgline and Smaïli (2021) developed an Amharic-English parallel corpus for both statistical machine translation and neural machine translation. In addition, Hadgu et al. (2021) introduced Lesan, a machine translation system for low-resource languages, built on the Transformer architecture, and provided an evaluation dataset. In a more recent effort, Biadgline and Smaïli (2021) expanded the Amharic-English corpus and experimented with tri-gram and four-gram SMT models, as well as GRU-based neural machine translation models. Despite these efforts, no prior studies have systematically analyzed the strengths and limitations of large language models in the context of Amharic machine translation.

In this study, we aim to identify the shortcomings of multilingual large language models in translating low-resource languages, using Amharic a case study. Using both human and automatic evaluation techniques, we identify common errors made by mLLMs in Amharic-English translations, providing insights for future researchers to improve Amharic machine translation. Specifically, we focus on the bidirectional Amharic-English MT system, with the following key objectives:

- To evaluate the performance of multilingual large language models in Amharic-English bidirectional machine translation.
- To identify the most common errors in Amharic-English bidirectional translations.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 19758–19768
November 12-16, 2024

- To provide insights into the limitations of mLLMs for low-resource languages like Amharic.

2 Related Works

This section reviews the related work to multilingual large language models for machine translations, with a particular focus on the evaluation for bidirectional Amharic-English machine translation.

Multilingual large language models are advanced neural architectures trained on large amounts of textual data, that can understand and generate multiple human languages. These models can be further fine-tuned for various downstream tasks such as machine translation, text classification, and question answering (Cooley and Tukey, 1965; Al-Khalifa et al., 2024). While the application of mLLMs in machine translation has gained significant attention in recent years, there is a limited research for evaluating the quality of LLMs in translation low-resource languages.

Recent Advancements in mLLMs , researchers have explored various techniques to improve the peformance of mLLMs. For instance, Hendy et al. (2023) conducted a comprehensive evaluation study for GPT models and analyzed different aspects such as their performance compared to state-of-the-art research and commercial systems. Their findings indicated that while GPT models provide competitive translation quality for languages with rich resources, their effectiveness is notably when applied to languages with limited resources.

On the other hand, significant advancements have been made by major companies, such as Meta, in developing LLMs capable of processing multiple languages, including those with limited resources. For instance, Fan et al. (2020) proposed a many-to-many multilingual translation model that facilitates direct translation between any pair of 100 languages. This model uses a data mining strategy based on language similarity to create a comprehensive multilingual dataset. Furthermore, the authors employed a Back-Translation technique to enhance the model's performance for zero-shot and low-resource language pairs, resulting in a training corpus of 7.5 billion sentences across 100 languages, including Amharic.

The "No Language Left Behind" Project is a notable effort by Meta AI, UC Berkeley, and Johns Hopkins University (Team et al., 2022). This interdisciplinary project aims to develop human-centered machine translation solutions for over 200 low-resource languages, including Amharic. Furthermore, the project focuses on dataset construction and model development to reduce the translation quality gap between resource-rich and resource-limited languages.

Evaluation of Machine Translation for Semitic Languages Despite the growing interest in applying LLMs to machine translation, there has been limited research evaluating their efficacy in this domain. A notable exception is the work by Abdelkadir et al. (2023), who assessed state-of-the-art MT for Tigrinya, a Semitic language closely related to Amharic and spoken by over 10 million people in Ethiopia and Eritrea. Their study utilized evaluation datasets spanning diverse domains, including Arts and Culture, Business and Economics, Politics, and Science and Technology. Employing the MQM-DQF error topology for translation assessment, the researchers concluded that mistranslation and omission were the most prevalent errors. This study provides valuable insights into the challenges of MT for Semitic languages, which may have implications for Amharic translation as well.

3 Experiment

We conducted our experiments to answer the following research questions regarding the peformance of multingual language models in low-resource languages, using Amharic as a case study:

RQ$_1$: *How do mLLMs compare in their performance in translation low-resource language when using automatic evaluation metrics?*

RQ$_2$: *What are the most frequent error types of mLLMs in the machine translation for low-resource languages?*

RQ$_3$: *Which mLLM performs better in translating low-resource language when using human evaluation?*

3.1 Setup

We perform our experiments to investigates the performance of mLLMs for low-resource language machine translation. First, we identified models trained on Amharic data. Then, we

Model	Params	Languages
NLLB-200	3.3B	200
NLLB-200	1.3B	200
NLLB-200-Distilled	1.3B	200
NLLB-200-Distilled	600M	200
M2Mm21.2B	1.2B	100
M2M418	418M	100

Table 1: Summary of multilingual Machine Translation models for Amharic–English translation

searched for evaluation datasets in both directions (i.e., from Amharic to another language and vice versa). The selected datasets were translated using the mLLMs, and the translations were evaluated using both automatic and human evaluation metrics. For human evaluation, volunteers were recruited as annotators, and they provided informed consent, ensuring transparency and adherence to ethical standards. To ensure the quality of the output data, inter-annotator agreement was calculated. After that, we compute the overall quality score of the annotated data. Figure 1 depicts our the pipeline of our approach for evaluating mLLMs in machine translation.

3.2 Language Model Selection

To identify mLLMs trained on Amharic data, we searched for available models on Google Scholar and Hugging Face. Among the available options, we found two notable models: NLLB200 and M2M100, both ared trained on multiple languages and supports Amharic language. The NLLB200 model is distributed under the CC-BY-NC license, while the M2M100 model is available under the MIT license. Table 1 provides an overview of these mLLMs, including number of parameters and the range of languages they support.

3.3 Dataset Selection

We followed the same approach for selecting the datasets as with the langauge models. This allowed us to find a dataset prepared by (Hadgu et al., 2021). This dataset includes 987 Amharic sentences and 1,915 English sentences, collected from various sources such as news article, Wikipedia, Twitter, and conversational sentences. The dataset is also publicly available under the CC BY 4.0 license.

3.4 Metrics

There are two primary methods for evaluating machine translation: automatic and human evaluation. Automatic evaluation relies on algorithms (e.g., BLEU, METEOR, and TER) to measure translation quality, while human evaluation involves professional translators to assess the accuracy and fluency of machine generated translations (Chatzikoumi, 2020).

3.4.1 Automatic Evaluation

For the automatic evaluation, we employed four metrics: BLEU (Bilingual Evaluation Understudy) (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ChrF++ (Banerjee and Lavie, 2005), and TER (Translation Edit Rate) (Snover et al., 2009) automatic evaluation metrics.

- BLEU measures the similarity between machine-translated text and reference translations by comparing overlapping n-grams.
- METEOR goes beyond BLEU by considering unigram precision and recall, as well as fragmentation, which accounts for word order and synonymy.
- ChrF++ evaluates both character and word-level n-grams, producing an F-score that reflects precision and recall, making it well-suited for morphologically rich languages like Amharic.
- TER assesses translation accuracy by calculating the number of edits (insertions, deletions, substitutions, and shifts) needed to make the machine-generated translation match the reference.

3.4.2 Human Evaluation

An initial review of the M2M model translations revealed that most results were inadequate, with many translations being unrelated to the source text. Given this poor performance, conducting a detailed human evaluation of M2M models was considered unnecessary due to the complexity and cost. For the error analysis, we adapted six error types from the Multidimensional Quality Metrics (MQM) framework: *Mistranslation*, *Addition*, *Omission*, *Grammar*, *Spelling*, and *Punctuation*. We also introduced an additional error type, Untranslated, to account for untranslated segments. The default severity levels*Neutral*, *Minor*, *Major*, and Criticalwere also adopted from the MQM guidelines. Below is a description of each error type and severity level as defined by MQM:[1]

[1] https://themqm.org/the-mqm-typology/

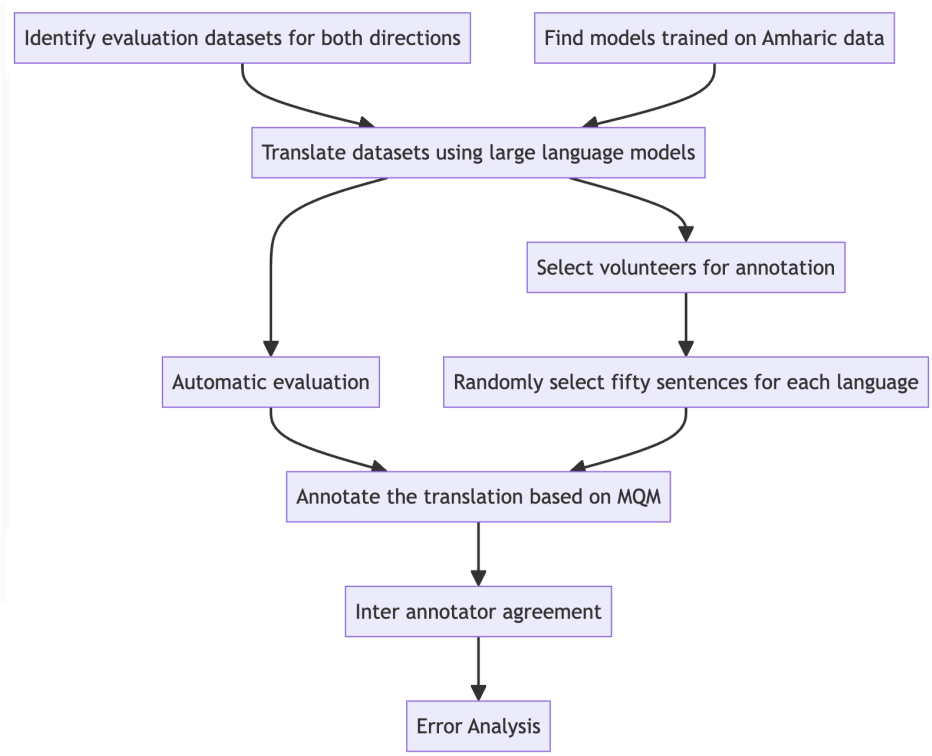

Figure 1: Our approach for Error Analysis in Multilingual Language Models

- **Mistranslation:** An error occurs when the target content does not accurately represent the source content.
- **Addition:** An error arises when the target content includes information that is not present in the source.
- **Omission:** An error occurs when content present in the source is missing from the target.
- **Untranslated:** An error arises when a segment intended for translation is omitted from the target content.
- **Grammar:** An error occurs when a sentence, phrase, or other text string violates the grammatical rules of the target language.
- **Punctuation:** An error occurs when punctuation does not conform to the conventions of the target language.
- **Spelling:** An error occurs when a word is misspelled.
- **Severity Levels:**
  - **1. Neutral Severity Level:** An acceptable translation error that does not significantly affect meaning.
  - **2. Minor Severity Level:** An error that has a minor impact on the accuracy and fluency of the translation.
  - **3. Major Severity Level:** An error is classified as Major if it affects the understandability, reliability, or usability of the content for its intended purpose.
  - **4. Critical Severity Level:** An error is classified as Critical if the translation is unfit for its intended purpose or poses a risk of serious physical, financial, or reputational harm.

We selected three annotators who are third-year university students from Addis Ababa University in Ethiopia. Thees students were chosen based on their proficiency in both English and Amharic. We conducted a 45-minute briefing session with the annotators to explain the dataset and the error annotation process using the MQM framework. We also provide them with detailed guidelines for completing the annotation form. Each student received compensation of $3,000$ Ethiopian birr (approximately $48.89).

Models	BLEU	chrf++	TER
NLLB3.3B	**26.7**	0.23	**0.36**
NLLB1.3B	21.01	**0.4**	0.32
NLLB1.3BD	15.36	0.28	0.3
NLLB600MD	12.84	0.3	0.13
M2M1.2B	13	0.31	0.9
M2M1418M	3.07	0.2	0.13

Table 2: Amharic to English automatic evaluation, answering $\mathbf{RQ}_1$

Models	BLEU	chrf++	TER
NLLB3.3B	22.00	**0.36**	0.86
NLLB1.3B	**23.03**	**0.36**	**0.86**
NLLB1.3BD	21.54	**0.36**	0.80
NLLB600MD	20.85	0.35	0.84
M2M1.2B	16.49	0.19	0.90
M2M418M	11.93	0.15	0.90

Table 3: English to Amharic automatic evaluation, answering $\mathbf{RQ}_1$

Following the briefing, the annotators filled out the form by first identifying errors in the translation and indicating their severity. In addition, they assessed the quality of each sentence, assigning a grade from 0 to 5; with 0 indicating a translation that was completely unrelated to the source and 5 is the almost perfect translation. Finally, we calculated the quality of each model and the inter-annotator agreement.

4 Result

4.1 Automatic Evaluation

Tables 3 and 2 show the results of automatic evaluations for the selected multilingual machine translation models translating between Amharic and English. The tables show that both NLLB3.3B and NLLB3.1B outperform the other models in all of the metrics in both Amharic to English and English to Amharic machine translation (answering $\mathbf{RQ}_1$). However, the accuracy of both M2M models is very low. In the first case, M2M418 scored only 3.07 BLEU, which was improved in English to Amharic translation to 11.93 BLEU.

4.2 Human Evaluation

For human evaluation, we did sentence- and word-level evaluation. For sentence-level evaluation, the annotators gave a score for each translated sentence a rating from 0 to 5. Then each score was converted to a percentage (see Table 4). The inter-annotator agreement was then calculated using Fleiss' Kappa. Table 5 shows Fleiss' Kappa coefficient. Fleiss' Kappa is a well-known measure

Models	Am-En	En-Am
NLLB3.13B	**63.6**	**76.1**
NLLB1.3B	28.67	42.67
NLLB1.3D	31	24.33
NLLB600M	37.47	38.6

Table 4: Sentence-level human evaluation of NLLB models in percent ($\mathbf{RQ}_3$)

Models	Am-En	En-Am
NLLB3.3	0.24	0.22
NLLB3.1	0.23	**0.40**
NLLB3.1D	0.23	0.21
NLLB600	**0.6**	0.19

Table 5: Fleiss' Kappa Coefficients for Am-En and En-Am

for inter-rater reliability (Moons and Vandervieren, 2023). In our cases, most of the coefficient values are more than or equal to 0.21. That means the annotators have a fair agreement (0.21 0.40); considering the complexity and nature of MT, this is an outstanding result. To calculate Fleiss' Kappa coefficient, we used a Python library called statsmodels.[2]

To compute word-level evaluation, we calculate the overall quality score (OQS) (Lommel et al., 2024) as follows: we define penalty rates for each error severity type; 0 for Minor, 1 for Major, 3 for Major, and 9 for Critical. Then OQS is calculated as:

$$\text{OQS} = (1 - (\text{PWPT} \times \text{PS})) \times \text{MSV} \quad (1)$$

Where PWPT (per-word penalty total) is calculated by dividing APT (Absolute Penalty Total) by EWC (Evaluation Word Count), which is the number of words in the source text:

$$\text{PWPT} = \frac{\text{APT}}{\text{EWC}} \quad (2)$$

APT is calculated by summing all the Error Type Penalty Totals (ETPTs). Each ETPT is the product of the number of errors for each error type and the severity penalty score for that error type:

$$\text{APT} = \sum(\text{ETPT}) \quad (3)$$

$$\text{ETPT} = \text{Error Count} \times \text{Penalty Rate} \quad (4)$$

[2] https://www.statsmodels.org/dev/generated/statsmodels.stats.inter_rater.fleiss_kappa.html

Models	Am-En	En-Am
NLLB3.3	**52.47**	**84.77**
NLLB1.3	27.23	75.3
NLLB1.3D	39.85	74.53
NLLB600D	36.7	80.03

Table 6: Overall Quality Score of NLLB Models ($\mathbf{RQ}_3$)

PS (Penalty Scaler) is a value from 0 to 1, with the default value being 1. For this experiment, we used PS value of 1. MSV (Maximum Score Value) represents a perfect upper score on a scale, with the default being 100. For this experiment, we used the default value.

As seen from 4 and 6, the NLLB3.3 model outperforms the rest of the models in sentence level and word level evaluation (Answering research question $\mathbf{RQ}_3$). In contrast to the automatic evaluation, NLLB600D performed better in human evaluation.

5 Error Analysis

In this section, we analyze the most frequent errors we observed in our experiment and their patters. Figures 2 and 3 show the the error distribution of the four models on the Lesan dataset. In this experiment, we saw that the primary cause of Mistranslation errors is the misinterpretation of personal and place names. For example, all models failed to correctly translate the name of Dr. Abiy, Ethiopia's prime minister, from Amharic to English. The M2M 1.2B model translated his name as "Al-Qaeda." Similarly, the name "Gojam," an area in the Amhara regional state of Ethiopia, where the regional capital (Bahirdar) located, was translated to "Gujarat" a city in Indiaby NLLB600. Additionally, the city name (seladinigayi) was inaccurately translated to "Saturday." However, the models successfully translated well-known place names such as Russia, Africa, and Pakistan. Our experiment shows that Mistranslation of entities are better handled in from English to Amharic than vise versa.

In some cases, models made critical Mistranslation errors. For instance, NLLB600 mistranslated the sentence "hahaha ire benatishi koka neberi yeteshegewi yih lewitima sewini hulu wiha kelemina tikuri kokani yawi newi" which roughly translates to "Haha, please, I thought it was coca; this will make everyone believe the color of water and coca is the same," to "Ha ha ha, your dad was a **cocaine smuggler**. This change makes everyone think that the color of water and black coca cola are the same," an entirely incorrect translation that could lead to serious legal issues.

Word	Translation	Correct
አቶ ወንድምኩን (ātō weni-dimikuni)	Mr. Brotherkun	Mr. Wendimkun
ኮካ ኮላ (koka kola)	KOKO KOLLA	Coca-cola
ሄንከን (hēnikeni)	crack cocaine	Heineken
ዋልያዎቹ (waliya-wochu)	guards	The Walias
ሰላድንጋይ (sela-dinigayi)	Salad Stone	Sela Dingay
ሰላድንጋይ (sela-dinigayi)	Saturday	Sela Dingay

Table 7: Examples of Mistranslation from Amharic to English

In addition, models tend to translate without considering context. For example, the models translated the sentence "እዝራ ፓውንድ የአሜሪካ ጸሐፊ ነበር" (izira pawinidi yeāmērika tseāfi neberi) to "Ezra Pound was an American secretary of state." The Amharic word "ጸሐፊ"(tsʾeāfi) was mistranslated directly to "Secretary," although it can mean both "Secretary" or "Writer" depending on the context. Here it should be translated to "Writer". Another example is the translation of "ጠቅላይ ሚኒስቴር" (tʾekʾilayi mīnīsitēri) where the NLLB3.3 model translated it to "general Minister" whereas the correct translation is "prime minister." The model ignored the context, which required the translation to be "Prime minister". Additionally, the word "Groundbreaking" was mistranslated as "የመሬት መሰራጨት"(yemereti meserachʾeti") unrelated to the intended meaning. The model interpreted "Ground" as "መሬት," referring to the earth's solid surface rather than its figurative sense.

In addition to translating words out of the context, the models tends to translate words literally. For instance, NLLB3.3 translated the company name "Friendly" to "wedajineti", which is incorrect; the name should either be left untranslated or transcribed into the Amharic alphabet. Similarly, the film title "The Sound of Music" was translated as "የሙዚቃው ድምፅ" (yemuzīkʾawi dimitsʾi) a direct translation of each word in the title.

Another common mistranslation error in-

Word	Mistranslation	Correction
The Sound of Music	የሙዚቃው ድምፅ (yemuzik'awi dimits'i)	ዘ ሳውንድ ኦፍ ሚውዚክ (ze sawinidi ofi mīwizīki) (Direct translation)
Smog	ሽንት (shiniti) (Urine)	ጭጋግ (ch'igagi)
Friendly	የወዳጅነት (yewedajineti) (Direct transaltion)	ፍሬንድሊ (firēnidlī)

Table 8: Examples of mistranslation errors from English to Amharic

volves technical and rare words. For example, the word "አኖዝሚያ" (ānozimīya) meaning "Anosmia," a medical term for partial or complete loss of smell, was completely mistranslated by all models to unrelated terms.

Similar to mistranslations, models frequently leave entities untranslated mostly in Amharic to English translations. Longer sentences often contain untranslated words. For instance the sentence "የኦሌ በቋሚነት መሾም ያልተዋጠላቸው አንዳንድ የኢትዮጵያ ማን ዮናይትድ ደጋፊዎች ተቃውሟቸውን ለመግለጽ ወደ ዲኤስቲቪ ቤት የሚወስዱ መንገዶችን በመዝጋት ላይ ናቸው" ("ye'olē bek'wamīneti meshomi yalitewat'elachewi ānidanidi ye'itiyop'iya mani yonayitidi degafīwochi tek'awimwachewini lemegilets'i wede dī'esitīvī bēti yemīwesidu menigedochini bemezigati layi nachewi") left untranslated it should be translated to "Ethiopian Manchester United Fans who do not support Ole's permanent appointment are closing the road to DSTV houses to express their anger." Another good example "የኢትዮጵያ ዜና አገልግሎት (ኢዜአ) እንደገና ሊቋቋም ነው" ("ye'itiyop'iya zēna āgeligiloti (īzē'a) inidegena līk'wak'wami newi") also left untranslated which should be translated to Ethiopian News Agency (ENA) is restablishing.

"Addition" errors occur when extra words are added in the translation. One of the primary reasons for "Addition" errors in our data is the repetition of words. For example, the phrase "ለምሳሌ የሞስኮው ፓፒሪ ተብሎ በሚታወቀው ክርስቶስ ከመወለዱ 1820 አመታት በፊት የተጻፈው የግብጻውያን መዝገብ ላይ፣ የጥረዛ ካልኩለስ ጭላንጭልን እናገኛለን።" (lemisalē yemosikowi papīrī tebilo bemītawek'ewi kirisitosi kemeweledu 1820 āmetati befīti yetets'afewi yegibits'awiyani mezigebi layi, yet'ireza kalikulesi ch'ilanich'ilini inagenyaleni) is translated as "For example, the Egyptian records of the birth of Christ, written in

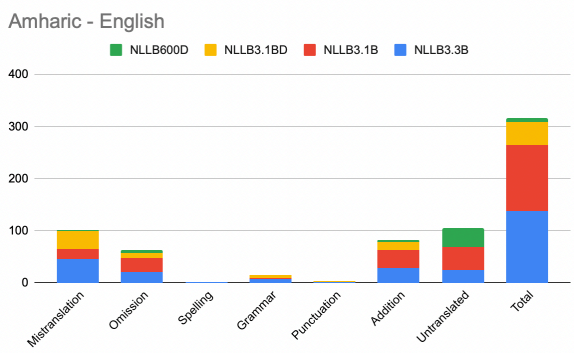

Figure 2: Error Distribution in Amharic to English Translation (Answering research question $\mathbf{RQ}_2$)

1820 years before the birth of Christ, known as the Moscow Papyri, contain the skulls of the Tzotzis' Calculus." Here, "The birth of Christ" is repeated twice, altering the meaning of the source sentence. For omissions, Entities are frequently omitted from translations, including names and parts of longer sentences. For instance, "*ለማንኛውም DSTV ያለው ሳይሆን ኖሮት የብሄራዊ ቡድኑን ጨዋታ ለማሳየት ፈቃደኛ የሆነ ቦታ ተለጥፎ ያያችሁ ወይም እዚህ ቢታይ የምትሉበት ቦታ ካለ ጠቁሙኝ ለማለት ነው። ከመዚዚሬ በፊት ኮምፓሴን ላስተካክል ብዬ ነው" ("lemaninyawimi DSTV yalewi sayihoni noroti yebihērawī budinuni ch'ewata lemasayeti fek'adenya yehone bota telet'ifo yayachihu weyimi izīhi bītayi yemitilubeti bota kale t'ek'umunyi lemaleti newi. kemezwazwarē befīti komipasēni lasitekakili biyē newi") was translated to "*If you have a place where you can watch the national team game, but not for any DSTV, or if you have a place where you can watch it, please tell me." Here, much of the text from the source is omitted. From grammar point of view Word order and subject-verb agreements are the main types of errors. As shown in Figures 2 and 3, there were very few errors related to punctuation and spelling.

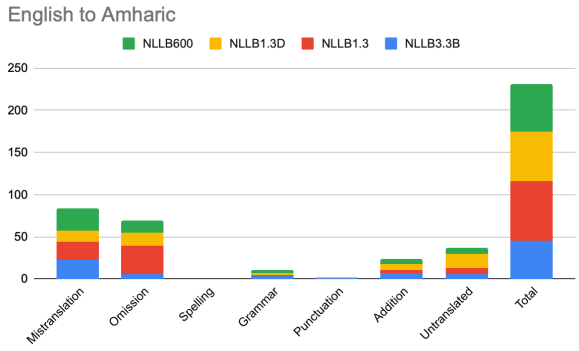

Figure 3: Error distribution in English to Amharic translation (Answering research question $\mathbf{RQ}_2$)

6 Conclusion

This paper evaluates the effectiveness of two public multilingual language models trained on Amharic: the M2M and NLLB models, both developed and released by Meta. In our evaluation, we we used a dataset created by Lesan AI. [3] First,we conducted our experiments with automatic evaluation, which show that the M2M models were ineffective in translating between both language directions. A preliminary review of the translations indicated that nearly all outputs were inadequate, suggesting that human evaluation of the M2M models may be unnecessary due to the complexity and associated costs.

In the human evaluation phase, we conducted two types of experiments: i) sentence level and ii) word level. Our findings indicate that across all evaluation techniques -- automatic evaluation and both sentence level and word-level human assessments-- the NLLB3.3B model outperforms the other models. Notably, while the NLLB600D model shows a superior performance in translating from English to Amharic compared to the NLLB1.3D and NLLB1.3 models, it performed worse in automatic evaluations. In the word-level human evaluation, the most common error types are: Mistranslation, Omission, and Untranslated segments with entities and longer sentences being the main reason for the errors. Grammar, punctuation and spelling errors rarely appear in our experiments. As part of our efforts to enhance translation quality, we released the annotated dataset under a permissive license. We hope that this research work will be a basis for future improvements of mLLMs for under resourced languages such as Amharic.

[3] https://github.com/dice-group/Error_Analysis_of_Multilingual_Language_Models_in_Machine_Translation

7 Limitations

Our study is constrained by the costs and time associated with human evaluation, leading to a focus on identifying the optimal models and the most common errors based on a single evaluation dataset. Ideally, a comprehensive analysis would include multiple datasets across diverse domains, such as medicine, technology, and law.

Acknowledgement

This work has been supported by the German Federal Ministry of Education and Research (BMBF) within the projects, COLIDE (grant no 01I521005D), KIAM (grant no 02L19C115), the European Union's Horizon Europe research and innovation programme (grant No 101070305), and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): TRR 318/1 2021 – 438445824.

References

Nuredin Ali Abdelkadir, Negasi Haile Abadi, and Asmelash Teka Hadgu. 2023. ERROR ANALYSIS OF TIGRINYA – ENGLISH MACHINE TRANSLATION SYSTEMS. In 4th Workshop on African Natural Language Processing.

Hend Al-Khalifa, Khaloud Al-Khalefah, and Hesham Haroon. 2024. Error analysis of pre-trained language models (plms) in english-to-arabic machine translation. Human-Centric Intelligent Systems.

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.

Yohanens Biadgline and Kamel Smaïli. 2021. Parallel Corpora Preparation for English-Amharic Machine Translation. In IWANN 2021 - International Work on Artificial Neural Networks, Conference Springer LNCS proceedings, Online, Spain.

Eirini Chatzikoumi. 2020. How to evaluate machine translation: A review of automated and human metrics. Natural Language Engineering, 26(2):137–161.

James W. Cooley and John W. Tukey. 1965. An algorithm for the machine calculation of complex Fourier series. Mathematics of Computation, 19(90):297--301.

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. CoRR, abs/2010.11125.

Andargachew Mekonnen Gezmu, Andreas Nürnberger, and Tesfaye Bayu Bati. 2021. Extended parallel corpus for amharic-english machine translation. CoRR, abs/2104.03543.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. Computational Linguistics, 48(3):673--732.

Asmelash Teka Hadgu, Abel Aregawi, and Adam Beaudoin. 2021. Lesan - machine translation for low resource languages. CoRR, abs/2112.08191.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. Preprint, arXiv:2302.09210.

Arle Lommel, Serge Gladkoff, Alan Melby, Sue Ellen Wright, Ingemar Strandvik, Katerina Gasova, Angelika Vaasa, Andy Benzo, Romina Marazzato Sparano, Monica Foresi, Johani Innis, Lifeng Han, and Goran Nenadic. 2024. The multi-range theory of translation quality measurement: Mqm scoring models and statistical quality control. Preprint, arXiv:2405.16969.

Filip Moons and Ellen Vandervieren. 2023. Measuring agreement among several raters classifying subjects into one-or-more (hierarchical) nominal categories. a generalisation of fleiss' kappa. Preprint, arXiv:2303.12502.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pages 311--318.

Matthew G Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. 2009. Ter-plus: paraphrase, semantic, and alignment enhancements to translation edit rate. Machine Translation, 23(2):117--127.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. Preprint, arXiv:2207.04672.

Senait Gebremichael Tesfagergish, Robertas Damaševičius, and Jurgita Kapočiūtė-Dzikienė. 2022. Deep learning-based sentiment classification of social network texts in amharic language. In ICT Innovations 2022. Reshaping the Future Towards a New Normal, pages 63--75, Cham. Springer Nature Switzerland.

Mulu Gebreegziabher Teshome, Laurent Besacier, Girma Taye, and Dereje Teferi. 2015. Phoneme-based english-amharic statistical machine translation. In AFRICON 2015, pages 1--5. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc.

A Appendix

A.1 Guideline for Filling the Survey

A.1.1 Introduction

This survey aims to identify common errors in the current multilingual machine translation models on Amharic--English translation. In the Excel document, fifty sentences are translated from Amharic to English or English to Amharic. The first column contains the original sentences, and the second shows its translation. Your task is to identify error types and their severity from the translation. The errors are mistranslation, addition, omission, untranslated, grammar,

punctuation, and spelling, and the severity is classified as neutral, minor, major, and critical. The following two sections explain how to identify the error types and their severity.

A.1.2 Error Types

Mistranslation An error occurs when the target content does not accurately represent the source content.
Examples: A source text states that a medicine should not be administered in doses greater than 200 mg, but the translation states that it should be administered in doses greater than 200 mg (i.e., negation has been omitted).

Addition Errors occur in the target content, including content not present in the source.
Examples: A translation includes portions of another translation that were inadvertently pasted into the document.

Omission An error occurs where content present in the source is missing from the target.
Examples: A paragraph present in the source is missing in the translation.

Untranslated An error occurs when a text segment intended for translation is omitted from the target content.
Examples: A sentence in a Japanese document translated into English is left in Japanese.

Grammar The error occurs when a text string (sentence, phrase, other) in the translation violates the grammatical rules of the target language.
Examples: An English text reads, ``The man was seeing his wife."

Punctuation Punctuation is incorrect according to target language conventions.
Examples: 1) An English text uses a semicolon where a comma should be used. 2) A two-digit year reference begins with an open single quote instead of a closed single quote (apostrophe). 3) An Amharic text uses a question mark instead of the anticipated semicolon to express a question.

Spelling Error occurs when a word is misspelled.
Examples: The German word ``Zustellung" is spelled ``Zustetlugn".

A.1.3 Severity Levels

1. Neutral Severity Level: The Severity Level of an error that differs from a quality evaluator's preferential translation or that is flagged for the translator's attention but is an acceptable translation.

2. Minor Severity Level: The Severity Level of an error that does not seriously impede the usability, understandability, or reliability of the content for its intended purpose, but has a limited impact on, for example, accuracy, stylistic quality, consistency, fluency, clarity, or general appeal of the content.

3. Major Severity Level: The Severity Level of an error that seriously affects the understandability, reliability, or usability of the content for its intended purpose or hinders the proper use of the product or service due to a significant loss or change in meaning or because the error appears in a highly visible or essential part of the content.

4. Critical Severity Level: The Severity Level of an error that renders the entire content unfit for purpose or poses the risk of serious physical, financial, or reputational harm. A single Critical Error would automatically trigger a Fail Rating in many systems.

A.1.4 How to Complete the Survey

In the second sheet of the Excel file, you will find a 6-column table. When you identify an error, write the number of the line in the "Line number" column, the error type in the "Error Types" column, and the severity in the "Severity" column as per the above instructions. Put the problematic word in the "Word" column, and if it is possible, explain the error in a few words and how it can be corrected in the "Explanation" column. In the last column named "Score", give a value for the translation

from 0 to 5. For example, if the translation is almost human-like, give it 5; if it is inadequate and not related to the source, give it 0.