A Human Perspective on GPT-4 Translations: Analysing Faroese to English News and Blog Text Translations

Annika Simonsen and Hafsteinn Einarsson
University of Iceland
Sæmundargata 2, 102 Reykjavík, Iceland
{annika, hafsteinne}@hi.is

Abstract

This study investigates the potential of Generative Pre-trained Transformer models, specifically GPT-4, to generate machine translation resources for the low-resource language, Faroese. Given the scarcity of high-quality, human-translated data for such languages, Large Language Models' capabilities to produce native-sounding text offer a practical solution. This approach is particularly valuable for generating paired translation examples where one is in natural, authentic Faroese as opposed to traditional approaches that went from English to Faroese, addressing a common limitation in such approaches. By creating such a synthetic parallel dataset and evaluating it through the Multidimensional Quality Metrics framework, this research assesses the translation quality offered by GPT-4. The findings reveal GPT-4's strengths in general translation tasks, while also highlighting its limitations in capturing cultural nuances.

1 Introduction

In the past decade, the field of Natural Language Processing (NLP) has seen a dramatic shift with the introduction of the attention mechanism and Transformer models, profoundly influencing the domain of Machine Translation (MT) (Bahdanau et al., 2014; Vaswani et al., 2017). One of the foremost challenges in MT is the scarcity of high-quality, human-translated data for low-resource languages. However, Large Language Models (LLMs) such as the Generative Pre-trained Transformer (GPT) models may present a solution to this challenge. These models are trained on vast amounts of data and have an impressive ability to generate native-sounding text, which they do by adapting based on the context presented in their training material (*in-context learning*) (Brown et al., 2020). Transformer models, such as GPT, are trained on multilingual data and have zero-shot translation capabilities which enables them to translate low-resource languages as well as high-resource languages. Therefore, this shift towards in-context learning signifies a breakthrough in NLP, where human-quality translation pairs can be generated without the input of a human translator. This has the potential of lowering the cost of making such data and improving the scalability of such an operation, which is vital for making smaller and more cost-effective models for MT.

This shift is particularly evident in the realm of MT datasets, where the gap between high-resource and low-resource languages remains a critical challenge. In the past, common methods to synthesize data for MT datasets were based on backtranslation (Sennrich et al., 2016; Poncelas et al., 2018; Poncelas et al., 2019). However, the quality of GPT models indicate that it is a better choice for synthesizing MT datasets (Hendy et al., 2023; Lyu et al., 2023).

The release of GPT-4 in March 2023 marked a significant milestone, with Jiao et al. (2023)'s pilot study that demonstrated its enhanced translation abilities in languages including English, German, Romanian, and Chinese, where its MT performance was comparable with state-of-the-art Neural Machine Translation (NMT) models. These results served as a motivation to explore the poten-

Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1), pages 24–36
June 24-27, 2024 ©2024 European Association for Machine Translation

tial of building MT datasets for low-resource languages with GPT models. Unlike traditional translation software, GPT-4 was not trained with the explicit purpose of MT, and in fact, it is not clear to what extent it was trained in MT at all. Further research, like Yang and Nicolai's (2023), demonstrated the potential of applying GPT-generated synthetic data in the context of MT. Their models translated from German (a high-resource language) and Galician (a low-resource language). Their findings revealed that while models trained solely on natural data outperformed those trained solely on synthetic data, the best performing model was the one trained on a combination of both datasets. These findings encourage the validation of translation quality in GPT models for low-resourced languages such as Faroese.

The population of the Faroe Islands is approximately 54,500 people (Statistics Faroe Islands, 2024), with the large majority speaking Faroese as their L1. At this time, Faroese MT resources are lacking, and the existing resources are not sufficient for training high performing MT models (Simonsen et al., 2022). Focusing on Faroese, this paper explores GPT-4's effectiveness in translating from Faroese to English. The potential of the GPT models raises the pivotal question: can the creation of MT resources for the low-resource language, Faroese, be automated, specifically through the capabilities of advanced models like GPT-4? To investigate this, the following contribution is made:

- A synthetic parallel dataset of 5,408 Faroese to English sentence pairs translated by GPT-4[1,2].
- A sample of 850 Faroese to English sentence pairs human evaluated and annotated with error labels from the Multidimensional Quality Metrics (MQM) framework by a native speaker[3,4].
- An in-depth analysis of the types of reoccurring errors that GPT-4 makes when translating from Faroese to English.

These contributions provide a detailed human examination of GPT-4's proficiency in translating Faroese to English, expanding on the current understanding of GPT-4's translation capabilities of low-resource languages.

2 Previous work

2.1 Faroese Parallel Datasets

There has been some preliminary work done in Faroese MT, specifically within the domain of creating parallel training data. However, the state-of-the-art neural network MT models of today need vast amounts of training data, an obstacle that Faroese is still facing. The largest available parallel training data for Faroese is *Sprotin's parallel corpus*[5] which was published on GitHub in 2020 and contains over 100k sentences human-translated from English to Faroese. This initiative was part of an effort to encourage Google to include Faroese in the Google Translate application (Hvidfeldt, 2020). In response, Microsoft released a model trained on this dataset in their MT system called *Microsoft Translator*. An Icelandic NLP company, Miðeind, also released a Faroese MT model trained on the same data around the same time on their MT system called *Vélþýðing* (Símonarson et al., 2021). For both systems it was apparent that the model performance was not high, which was likely due to the small amount of training data. Faroese was never added to Google Translate and is also currently no longer supported on Vélþýðing. More recently, Meta launched the *No Language Left Behind* (NLLB) project which aims to bridge the gap in the performance between high- and low-resource languages in MT (Team et al., 2022). They published a series of open-sourced MT models called NLLB[6] and a human-translated parallel dataset called FLORES-200[7] which covers over 200 languages, including Faroese. The NLLB model's capability to translate Faroese appears promising based on preliminary experiments made

[1] https://huggingface.co/datasets/AnnikaSimonsen/GPT-4_FO-EN_parallel_news_sentences
[2] https://huggingface.co/datasets/AnnikaSimonsen/GPT-4_FO-EN_parallel_blog_sentences
[3] https://huggingface.co/datasets/AnnikaSimonsen/GPT-4_FO-EN_parallel_news_sentences_MQM
[4] https://huggingface.co/datasets/AnnikaSimonsen/GPT-4_FO-EN_parallel_blog_sentences_MQM
[5] https://raw.githubusercontent.com/Sprotin/translations/main/sentences_en-fo.strict.csv
[6] https://huggingface.co/facebook/nllb-200-distilled-1.3B
[7] https://huggingface.co/datasets/facebook/flores

by the authors of this paper, a notable achievement considering the majority of its training data for Faroese is not genuinely parallel but is instead incorrectly aligned Faroese to English data.

Building upon the foundation of leveraging linguistic relations for enhancing machine translation in low-resource languages, recent studies have begun exploring the potential of utilizing phylogenetic information from high-resource languages within the same language family. For example, Snæbjarnarson et al. (2023) demonstrated that by incorporating resources from closely related Scandinavian languages, the performance of NLP tasks in Faroese could be substantially improved. This method marks a departure from the traditional 'one-size-fits-all' approach taken by widely used multilingual transformers like mBERT or XLM-R, advocating instead for a tailored strategy that considers the unique linguistic heritage of each language family. Such insights reveal the advantages of a more focused approach in data augmentation and model training, particularly for languages like Faroese that have fewer resources. Additionally, in Scalvini and Debess' (2024) upcoming publication, they highlight the effectiveness of GPT-SW3, a Scandinavian-focused LLM, in leveraging the linguistic similarities between Faroese and its Nordic counterparts to enhance translation accuracy and facilitate data augmentation.

2.2 Generating synthetic parallel data using GPT models

As mentioned in the introduction, there have been recent studies that explored using generative LLMs like ChatGPT or GPT-4 to create synthetic parallel data for training MT models. Inspired by findings that GPT-4 could match the translation abilities of commercial NMT systems, Yang and Nicolai (2023) explored training translation models for German and Galician using ChatGPT-generated synthetic data. In their study, they compared models trained on natural data from TED Talks with those trained on synthetic data, created by translating seed words and sentences into English via ChatGPT. Although models trained on real data performed better than those trained on synthetic data, those trained on a mix of real and synthetic data (augmented model) showed improved translation quality for both languages. Interestingly, synthetic Galician data yielded better translation quality than the German synthetic data. However, the study showed that there was a lower linguistic diversity in the synthetic data compared to natural data, evidenced by a lower type-token ratio (TTR), indicating a repetition of sentences and limited vocabulary use. This highlights challenges in leveraging LLMs for low-resource synthetic data creation. Nonetheless, supplementing real data with synthetic data remains a promising strategy for training MT models for low-resource languages (Poncelas et al., 2018; Poncelas et al., 2019).

```
function_descriptions = [
    {
        "name": "translation_analysis",
        "description": "The function analysis text that
      has been translated from Faroese to English. The
      input translation should be of exceptionally high
      quality.",
         "parameters": {
             "type": "object",
             "properties": {
                 "sentence_analysis_list": {
                     "type": "array",
                     "items": {
                         "type": "object",
                         "properties": {
                             "original":
                             {"type": "string"},
                             "translation":
                             {"type": "string"},
                         }
                     }
                 }
             },
             "required": ["sentence_analysis_list"]
         }
    }
]
```

Listing 1: JSON schema for translation analysis.

3 Experimental setup

This section outlines the methodology used to examine GPT-4's effectiveness in generating parallel data for Faroese, a language with limited resources. Firstly, the experiment involves generating synthetic parallel data using GPT-4. This output was then evaluated using the Multidimensional Quality Metrics (MQM) framework with a single native speaker as an annotator.

3.1 Prompting Approach

To generate the synthetic parallel data, a structured prompting approach with GPT-4 was employed, setting the temperature parameter to 0 to guarantee uniform and deterministic output. The experiment extracted information from Faroese news and blog texts through OpenAI's API, organizing it according to a specific JSON format (as detailed in Listing 1). This format instructed GPT-4 to translate texts sentence by sentence.

3.2 Data Preparation

3.2.1 GPT-4 Parallel Sentences

The parallel sentences generated by GPT-4 were derived from the Basic Language Resource Kit for Faroese 1.0 text corpus (Simonsen et al., 2022). During the first round, news texts from the online newspapers *Dimmalætting* and *Portalurin* were processed, with GPT-4 translating each document sentence by sentence, yielding 3,735 Faroese to English news sentence pairs. Subsequently, blog texts were translated using the same procedure, including works from *Egið Rúm* by Marna Jacobsen[8], *BAVS* by Bergljót av Skarði[9], and *BirkBlog* by Birgir Kruse[10], resulting in 1,673 sentence pairs from blogs. The aim was to capture a diverse representation of GPT-4's translation skills by combining news and blog texts, acknowledging their distinct genres. In total, there were 5,408 generated sentence pairs.

3.3 Human Evaluation

A subset of the GPT-4 generated parallel data was sampled for human evaluation using the MQM framework[11]. MQM incorporates more than twenty traditional translation quality metrics and provides a detailed catalogue of over 100 potential issues for assessing translations and source texts. It is designed as a flexible master list from which specific issues can be chosen based on the translation task at hand, allowing for customization to meet diverse requirements. In the context of this study's MT evaluation, a tailored version of the MQM framework, as adapted by Freitag et al. (2021) was employed. An overview of the MQM error categories utilized by Freitag et al. (2021), along with their descriptions, is presented in Table 1.

The sample that was chosen for human evaluation was created by choosing articles randomly and then evaluating the chosen articles, sentence by sentence. There was only one annotator, author of this paper, who is a linguist and native speaker of Faroese. In total, 425 news sentence pairs and 425 blog sentences pairs were human evaluated. The evaluation was carried out in a Google Sheet spreadsheet (see Figure 1). To calculate the MQM score, the official MQM spreadsheet was used. This spreadsheet contains all relevant formulas to calculate the MQM score, also known as the *Overall Quality Score* (OQS).

4 Results

The results for the MQM evaluation is summarized in Table 2. Overall, the quality of translations is high as indicated by the MQM score or *Overall Quality Score*.

As seen in Table 2, the predominant severity level assigned for MQM was *minor*. This classification was used when translations were not technically accurate but still conveyed the intended meaning. The *major* category was designated for errors in translation that obscured or altered the meaning, while *critical* was reserved for when the translation got offensive or dangerously misinformed. Notably, major accuracy errors occurred more frequently in blogs than in news articles, often arising in idiomatic expressions and set phrases. In the case of news articles, there were three instances classified as *critical*[12]:

1) *Example sentence containing critical error from Portalurin article.*

- **FO**: *Somuleiðis skulu dagføringar gerast á Vágs høll við máling og wc til* ***rørðslutarna*** *skal gerast í ganginum millum VB húsið og Vágs Høll.*
- **ENG**: "Similarly, updates should be made to the Vágur hall with regards to painting and a toilet for **disabled** that should be made in the corridor between the VB house and Vágur Hall."
- **GPT-4:** "Similarly, updates should be made to Vágur hall with painting and a toilet for **the pipe players** should be made in the corridor between the VB house and the Vágur Hall."

The original sentence contains a spelling mistake; *rørðslutarna* is supposed to be spelled *rørslutarnað* which translates to "disabled". GPT-4 translated this term to "the pipe players".

[8]Jacobsen shares insights from her personal life, coupled with reviews of music, books and movies. Available at: https://marnakj.wordpress.com/.

[9]BAVS is centered on personal experiences, culture, and travel. Available at https://b-av-s.blogspot.com/

[10]A blog focusing on cultural events along with reviews of movies, music, and more. Available at: https://birkblog.blogspot.com/

[11]https://www.qt21.eu/

[12]The order in the sentence examples is as follows: **FO** (original Faroese sentence from dataset), **ENG** (a translation provided by an author of this paper) and **GPT-4** (GPT-4's translation of the original Faroese sentence).

Error Category	Description
Accuracy	
Addition	Translation includes information not present in the source.
Omission	Translation is missing content from the source.
Mistranslation	Translation does not accurately represent the source.
Untranslated text	Source text has been left untranslated.
Fluency	
Punctuation	Incorrect punctuation (for locale or style).
Spelling	Incorrect spelling or capitalization.
Grammar	Problems with grammar, other than orthography.
Register	Wrong grammatical register.
Inconsistency	Internal inconsistency (not related to terminology).
Character encoding	Characters are garbled due to incorrect encoding.
Terminology	
Inappropriate for context	Terminology is non-standard or does not fit context.
Inconsistent use	Terminology is used inconsistently.
Style	
Awkward	Translation has stylistic problems.
Other	Any other issues.
Source error	An error in the source text.
Non-translation	Impossible to reliably characterize the 5 most severe errors.

Table 1: Overview of MQM label hierarchy.

Faroese	English translation	Errors	Severity
Danska rokktrioin, sum legði ríkið fyri sínar føtur í 90´unum, eigur eitt heilt serligt plass í hjartanum á mongum føroyingi.	The Danish rock trio, which conquered the country in the 90s, holds a special place in the hearts of many Faroese.	style(føroyingi-Faroese/Faroe Islanders)	minor
Trioin gav út plátuna Dizzy Mizz Lizzy í 1994, og hon streyk upp á tónleikatindarnar alt fyri eitt.	The trio released the album Dizzy Mizz Lizzy in 1994, and it shot up the music charts immediately.		
Útgávan var serstøk, tí hon var á tremur við hittum.	The release was special because it was on par with others.	accuracy(á tremur við hittum-on par with others/full of hits)	major

Figure 1: Figure showing the human evaluation method for the GPT-4 generated parallel sentences in Google Sheets.

2) *Example sentence containing critical error from Dimmalætting article.*

- **FO:** *Orsøkin er, at svenski Umhvørvisflokkurin, ið var í stjórn saman við Sosialdemokratunum, hevur vent samgonguni bakið, eftir at tað gjørdist greitt, at* ***borgarliga andstøðan*** *fekk ein meiriluta við at atkvøða sína fíggjarlóg ígjøgnum.*
- **ENG:** "The reason is that the Swedish Environmental Party, which was in government with the Social Democrats, has turned its back on the coalition, after it became clear that **the civil opposition** got a majority by voting their budget through."
- **GPT-4:** "The reason is that the Swedish Environmental Party, which was in government with the Social Democrats, has turned its back on the coalition, after it became clear that **the bourgeois opposition** got a majority by voting their budget through."

The word for "civil opposition" has been translated into "bourgeois opposition", which could have negative connotations. The third example is in the same article where "civil budget" was translated into "bourgeois budget".

In the blog texts, a single critical error was identified in the sample, involving the mistranslation of the term *at ræsa* — the traditional Faroese method of fermenting meat through dry-aging. It was incorrectly translated as "raw". This misinterpretation could be seen as dangerously misleading and potentially harmful to someone's health, especially if the food had not been pre-cooked as could be inferred from the context given correct world-knowledge:

3) *Example sentence containing critical error from BirkBlog.*

- **FO:** *Eg vildi smakka* ***ræstu*** *pylsuna.*
- **ENG:** "I wanted to taste the **Faroese dry-aged** sausage."
- **GPT-4:** "I wanted to taste the **raw** hot dog."

The Overall Quality Score (OQS), as detailed in Table 2, serves as a metric for assessing translation quality. It is derived through a systematic procedure: annotators input error annotations into a matrix (see Figure 2), assigning them numerical values based on error type and severity, to obtain the Absolute Penalty Total (ABT). The OQS calculation incorporates several factors, including the Per-word Penalty Total, calculated by dividing the ABT by the total word count (EWC); the Overall Normed Penalty Total, which adjusts the per-word penalty in relation to the total number of reference

Category	News Sentences (425)			Blog Sentences (425)		
	Minor	Major	Critical	Minor	Major	Critical
Accuracy	43	26	1	31	54	1
Fluency	9	0	0	8	5	0
Terminology	76	14	2	13	8	0
Style	35	2	0	11	3	0
MQM Score			**94.41**			**88.38**

Table 2: MQM evaluation results for 425 news sentences and 425 blog sentences. The MQM score is also known as the Overall Quality Score (OQS). The weights are minor (-1), major (-5) and critical (-25). A higher score (with a maximum of 100) corresponds to better performance.

	A	B	C	D	E	F	G	H
1	MQM Scorecard: Top-Level Error Typology with 4 Severity Levels							
2								
3		*Error Severity Levels:*	Neutral	Minor	Major	Critical	Error Type Penalty Total	
4		*Severity Penalty Multipliers:*	0	1	5	25		
5	ET Nos	Error Types	Error Counts				ET Weights	ETPTs
6	1	Terminology	2	7	7	0	1.0	42.0
7	2	Accuracy	4	14	7	1	1.0	74.0
8	3	Linguistic conventions	1	23	9	0	1.0	68.0
9	4	Style	5	7	3	0	1.0	22.0
10	5	Locale convention	1	12	5	0	1.0	37.0
11	6	Audience appropriateness	0	2	1	0	1.0	7.0
12	7	Design and markup	0	6	1	0	1.0	11.0
13	8	Custom						
14							Absolute Penalty Total:	261.00
15								
16		Evaluation Word Count:	10184				Per-Word Penalty Total:	0.0256
17		Reference Word Count:	1000				Overall Normed Penalty Total:	25.63
18		Scaling Parameter (SP):	1.00				Overall Quality Score:	97.44
19		Max. Score Value:	100.00					
20		Threshold Value:	85.00				Pass/Fail Rating:	Pass

Figure 2: Figure showing the Overall Quality Score card from `https://themqm.org/`.

words; and the Overall Quality Fraction, achieved by dividing the ABT by the EWC. The final Overall Quality Score is computed by subtracting the result of multiplying the per-word penalty score by the highest possible score from 1, thereby converting the score into a more recognizable percentage format. This approach integrates a meticulous evaluation of translation inaccuracies with a comprehensive scoring framework to measure and express the quality of translations quantitatively.

In Table 2, the Overall Quality Score demonstrates high performance in translation quality for both text genres, with news articles achieving a score of 94.41/100 and blogs receiving 88.38/100. According to the MQM framework, a score within the range of $94 \leq x < 98$ signifies a high level of quality, whereas scores in the range of $80 \leq x < 94$ are indicative of a good quality level, as outlined in Talhadas (2023). Following this, a qualitative analysis is provided to examine the specific types of translation errors GPT-4 made while translating from Faroese to English.

4.1 Most Common FO–EN Translation Errors by GPT-4

There is a general pattern in the types of errors that GPT-4 makes when translating from Faroese to English. A prominent error involves the translation of the Faroese term for the Danish currency used in the Faroe Islands, *krónur*. GPT-4 often renders these as "crowns" or "kroner", whereas the conventional translation should be "DKK" or "Danish crowns." Additionally, we observed four times that *føroyingur* was translated to "the Faroese" in the sample, but a more precise translation would be "a Faroe Islander" or "a Faroese person." A check on the rest of the translated data revealed that this was a common mistranslation.

Another consistent error is the translation of *korona* to "corona." While not incorrect, "COVID-19" is the term more frequently used in English news articles, making it a more suitable translation in those contexts. Given that "COVID" was not adopted into Faroese during the pandemic, Faroese news texts use *korona* instead.

Subsequent sections will detail the other preva-

lent errors, categorized by their types, to provide a comprehensive overview of the translation challenges encountered. See Table 3 in the Appendix for a quantitative analysis of the errors analyzed in this section.

4.1.1 Named Entities (NEs)

GPT-4 did not consistently translate all NEs incorrectly, but did manage to correctly translate certain NEs, such as names of institutions, e.g. *Ráðið fyri Ferðslutrygd* ("Council for Traffic Safety"). It also frequently accurately converted people's names from the dative to the nominative case in English, such as translating *Mariu* to "Maria". Nevertheless, there are many examples of translation errors with NEs. For instance, the short form name *Setrið* for *Fróðskaparsetrið* was incorrectly translated literally as "Center" rather than "The University of the Faroe Islands" or simply "Setrið". Additionally, *Okkara Voxbotn* was inaccurately translated to "Our Voxbotn" instead of preserving its original name, which is associated with a brewery named *Okkara* that sponsors a music festival that takes place in the harbour named *Voksbotn*. These examples illustrate the nuanced difficulties GPT-4 faces with NEs in the context of Faroese to English translation.

4.1.2 Correct Translation, Wrong Terminology

While GPT-4 frequently chooses accurate translations for words, it often selects the wrong terms. For instance, in some contexts *skeið* is translated as "course" when it is supposed to be "workshop," and *eldraøki* is translated as "elderly area" instead of "elderly affairs". The term *øki* can be translated as "area" only when it is referring to a physical place. In this context, the term was used in a sentence from an article about a financial budget of a town, where the taxes had been increased to cover elderly affairs. Therefore, while these translations are technically correct, they are not entirely appropriate in these specific contexts.

4.1.3 Idioms and Fixed Phrases

GPT-4 often encounters difficulties with idiomatic expressions and fixed phrases, particularly in the Faroese blog texts. These phrases frequently undergo literal translation, which misses their nuanced meanings. For instance, the phrase *at fáa sær okkurt gott* is directly translated as "getting oneself something good" rather than capturing the intended meaning of "getting something to eat." Similarly, a well-known Faroese phrase, *er ikki sum at siga tað*, intended to convey that something is not easy, is translated by GPT-4 in a literal manner as "is not like saying it."

4.2 Icelandicisms

GPT-4 often confounds Faroese with Icelandic, likely due to the fact that GPT-4 has been trained on significantly more Icelandic data than Faroese. This leads to what we term "Icelandicism", where translations mistakenly apply Icelandic meanings. Examples include translating menning as "culture" instead of "progress", sætti as "sweet" instead of "sixth" and bleytur as "wet" rather than "soft". Here, it is presumed that the Faroese word sætti is confused with the Icelandic word sætur and the Faroese word bleytur is conflated with the Icelandic word blautur.

4.2.1 Cultural Context

GPT-4 often misses the cultural nuances in its translations, leading to misunderstandings of certain terms. For example, it interprets *ríkið* as "country" instead of "kingdom". In Faroese contexts, *ríkið* typically refers to the Danish Kingdom rather than the Faroe Islands. This misinterpretation might also reflect an Icelandicism. Other Faroese terms that are commonly mistranslated include *fiskaplassið*, which refers to a stone-paved area for drying fish but gets translated as "fish place," and *hoyggjhús*, which is translated as "living room" instead of "hay barn". Occasionally, these culturally specific terms are translated into nonsensical words. For instance, the Faroese term for "paternal granduncle", *abbabeiggi*, was erroneously translated as "abbess." The term *abbabeiggi* is culturally significant, as, although Icelandic also has a term for the brother of your grandfather, *afabróður*, it is not as commonly used as in Faroese. Notably, Danish lacks a term for this specific type of granduncle. Another example of a mistranslation is the Faroese word for a national dish, pilot whale steak (*grindabúffur*), which was translated to the nonsense word "grindabuffi".

This examination underscores that although GPT-4's FO–EN translations are of commendable quality, they exhibit specific and frequently recurring mistakes, notably in handling cultural subtleties and idiomatic expressions. Further qualita-

tive analyses of errors are deferred to the appendix.

5 Discussion

According to the human evaluation of the GPT-4 translation data, GPT-4 has demonstrated its proficiency in translating from Faroese to English, especially in the context of news articles. However, the translations are not perfect and we address the limitations later in this section. This finding of translation quality is consistent with recent research indicating GPT-4's effectiveness in translating from low-resource languages to English, as highlighted in studies by Bang et al. (2023), Jiao et al. (2023), and Yang and Nicolai (2023). The model's particular strength in news translation is likely due to its extensive training on a wide array of news texts, which is abundantly available online for collection. However, when it comes to blog texts, which are rich in idiomatic expressions and fixed phrases, GPT-4's performance dips slightly (from 94.41 to 88.38). This drop in performance suggests that while GPT-4 can generate high-quality translations, its capability diminishes with content that heavily features language-specific idioms and cultural nuances. Yet, the synthetic parallel sentences generated by GPT-4 present a "Silver Standard" resource for training MT models for Faroese, complementing the "Gold Standard" human-translated data. Although not a substitute for human translation, the combination of synthetic and human-generated data could potentially enhance the training materials available for Faroese MT models (for German and Galician, see Yang and Nicolai (2023); for English, German and Turkish, see Sennrich et al. (2016). However, to fully assess the impact of GPT-4 generated parallel data on MT model performance, a larger dataset would be ideal. For this study, the collection was limited to 5,408 sentence pairs due to cost considerations and the licensing restrictions imposed by OpenAI on their model's output[13].

During the study, a preliminary experiment was conducted to see how well GPT-4's translation performed from English into Faroese, which revealed significant limitations. The model often failed to construct grammatically correct Faroese sentences, frequently producing outputs that appeared to be an amalgamation of Icelandic and Faroese. This finding corroborates previous research indicating that GPT-4's capabilities in translating from English to low-resource languages remain constrained. Studies by Hendy et al. (2023), Lyu et al. (2023), Jiao et al. (2023), and Yang and Nicolai (2023) have similarly documented these challenges, reinforcing the observation that GPT-4's performance in such translation tasks is not yet satisfactory.

[13]At the time of usage, gpt-4-0613 cost $0.03 for every input token and $0.06 for every output token.

6 Limitations

6.1 GPT-4's Splitting of Sentences

GPT-4 did not consistently split the Faroese text into sentences although it was explicitly instructed to do so using our function-callin approach to extract output in a structured manner. An analysis of a random selection of 500 rows from the parallel dataset generated by GPT-4 revealed 29 cases of improper sentence division. This means that GPT-4 incorrectly split the Faroese sentences 5.8% of the time. This could also be related to the reason why GPT-4 struggled with translating NEs. NEs are notoriously difficult for MT models to handle accurately, and Named Entity Recognizers (NERs) are often employed alongside these models to enhance performance (Babych and Hartley, 2003). In the early stages of developing the GPT-4 parallel data for this experiment, attempts were made to have GPT-4 label the Faroese text with NE labels. However, these efforts were unsuccessful, leading to the exclusion of this step from the process. This difficulty likely stems from GPT-4's inadequate ability to recognize Faroese NEs, contributing to its struggles with their translation.

6.2 Systematic Translation Errors

While GPT-4 delivers translations of high quality from Faroese to English, it is worrying to see that the errors it makes are often specific to Faroese context and culture. These mistakes do not seem to be random but show a pattern that could negatively affect the efficacy of an MT model trained with such data. A possible remedy might have been to enrich GPT-4's contextual understanding, perhaps by feeding it Wikipedia articles that encapsulate key facts about Faroese culture and the Faroe Islands, or by exposing it to Faroese texts across different genres. This enhanced prompting strategy, not dissimilar to few-shot prompting (Brown et al., 2020) could have helped GPT-4 in situations where its grasp of context and global knowledge fell short. Ultimately, the synthetic parallel data

produced by GPT-4 ought to be considered as "Silver Standard", rather than "Gold Standard" data, which is typically human-translated. Drawing parallels to the findings of Yang and Nicolai (2023) regarding German and Galician ChatGPT-generated parallel data, it becomes apparent that Silver Standard Data holds value, particularly when combined with human-translated data for training Faroese MT models to maximize performance.

6.3 Lack of MQM annotators

It is crucial to acknowledge that since there was only one annotator, the MQM scores should not be compared with those from projects that had multiple annotators, under the assumption that the result for Faroese is not as robust as for other languages (i.e. due to potential individual biases). The primary aim of conducting the MQM evaluation was to delve into error analysis and obtain a comprehensive understanding of the translation quality. Additionally, the Faroese annotator chose not to apply the "neutral" error weight during the MQM assessment, a deviation from conventional practices. This decision was made because labeling an error as "neutral" seemed inappropriate when such a categorization is typically reserved for instances deemed not to be the translator's fault and, in this case, the translator is a language model. Looking back, this neutral category might have been applicable for source text errors, such as typos, but ultimately, these errors were given the label "source error", so the resulting score was not affected as "source errors" count the same as a "neutral" error. Only 12 source errors were found in total, and only four of them resulted in a translation error. Determining whether an error is attributable to the language model presents its own challenges. Furthermore, given that the MQM framework is designed for evaluating both human- and machine translation, applying it uniformly to both can be problematic. For instance, human translators often tailor their work to a client's specific style requirements, which can range from general and succinct to verbatim translations, depending on whether clarity or fidelity is prioritized. Current MT models, however, lack the capability to adjust their output based on stylistic preferences without specific training for each requirement. Nevertheless, LLMs like GPT-4 have shown the ability to adapt to given instructions, suggesting they can be directed to follow certain styles or formalities. Future research may need to reconsider how we evaluate LLMs like GPT-4, taking into account their unique capabilities and limitations.

6.4 OpenAI and Model Ownership

OpenAI's terms of use (OpenAI, 2023) stipulate that while users are granted ownership of the output generated by its services, there are restrictions when it comes to using GPT-4 output for model training. Specifically, users are prohibited from using the output to develop models that compete with OpenAI. As a result, the authors of this paper limited their generation to approximately five thousand parallel sentences with GPT-4, as they would not have been able to share any models fine-tuned using this training data. Similarly, AI META's Llama 2 model permits derivative works but forbids their use in enhancing language models other than Llama (Meta AI, 2023), which would have prevented the authors from sharing their fine-tuned MT models had they used Llama 2 instead of GPT-4.

However, there are open-source LLMs that serve as alternatives, some of which aim to address the lack of diversity in the text used to train LLMs. Notable efforts include AI Sweden's GPT-SW3 (Swedish Government, 2023), focused on Nordic languages, and the upcoming Horizon Europe funded TrustLLM, aiming for an open, trustworthy, and Germanic language-focused LLM[14]. AI Sweden offers a flexible license for GPT-SW3 (AI Sweden, 2023), exemplifying the push towards democratizing LLM access. It is worth noting that there are significant differences in parameter size between these models, with GPT-SW3's largest instruct model having 20B parameters, Llama 2's biggest instruct model having 70B parameters, and GPT-4 believed to have over a trillion parameters. Another recently published open model is Mistral's Mixtral 8x7B, which is under the Apache 2.0 license and is reported to either match or outperform Llama 2 and GPT-3.5 on most standard benchmarks (Mistral AI team, 2023). These open models provide potential alternatives for future work in automating Faroese NLP resource creation.

6.5 Future Work

This research has identified several promising directions for future work in Faroese MT. Firstly,

[14] https://trustllm.eu/

the potential of synthetic parallel data produced by LLMs like GPT-4 for Faroese remains largely unexplored. Future efforts should focus on creating a larger corpus of synthetic parallel sentences covering a wider range of text genres beyond news and blogs. This approach would provide insights into how effectively such data can train more robust MT systems. However, licensing restrictions associated with some LLMs may necessitate a shift towards openly available models, such as GPT-SW3, the forthcoming Germanic LLM from the TrustLLM project, or Mistral's Mixtral 8x7B. These open models would facilitate the generation of larger datasets and ensure the ability to freely share and distribute the resulting works, aligning with research efforts aimed at enhancing NLP capabilities for low-resource languages like Faroese. Scalvini and Debess (2024) have demonstrated the merits of using language-family-specific models, such as GPT-Sw3, in refining translation accuracy and facilitating data augmentation efforts for Faroese.

Secondly, there is currently no human-translated parallel dataset for Faroese derived from monolingual Faroese texts. Existing datasets, such as FLORES-200 and the Sprotin parallel corpus, are translations from English and do not accurately reflect Faroese-specific expressions and terminologies. Consequently, the synthetic parallel data generated by GPT-4 also falls short in capturing these unique Faroese nuances. Therefore, developing a human-translated parallel dataset centered around Faroese monolingual content, with an emphasis on capturing the richness of Faroese cultural and linguistic elements, would be highly advantageous for future research in Faroese MT.

Finally, recent advancements in models like Gemini 1.5 Pro, which can process exceptionally long contexts, have opened up new prospects for MT in Faroese. Gemini 1.5 Pro has demonstrated its ability to learn new languages from a minimal set of instructional materials. Specifically, with only 500 pages of linguistic documentation and approximately 400 parallel sentences, it managed to learn and translate from English to Kalamang, a critically low-resource language with minimal online presence, achieving translation quality comparable to human learners (Gemini Team, 2024). This success suggests that for Faroese, leveraging Faroese grammar books and lexical resources in the translation context could make high-quality translation not only feasible but also efficient. This promising approach warrants further investigation in future research.

7 Conclusion

In conclusion, this paper has demonstrated the potential of GPT models like GPT-4 in generating synthetic parallel data, potentially mitigating the scarcity of high-quality, human-translated datasets. Through a detailed analysis of GPT-4's translation from Faroese to English, including a synthetic parallel dataset and an MQM framework-based evaluation, we have uncovered both strengths and limitations of employing GPT models for MT. While GPT-4 shows promise in generating translations that could serve as valuable training data, challenges remain, particularly with translations that involve cultural and contextual nuances. This exploration not only contributes to the understanding of GPT models' capabilities in translating low-resource languages but also sets the stage for future research directions. By integrating synthetic and human-generated data, there's potential to enhance MT models for Faroese, pushing the boundaries of accessibility and quality in MT for low-resource languages. This study underscores the necessity for ongoing research to fully leverage the capabilities of advanced models like GPT-4, aiming for a future where no language is left behind in the digital age.

Acknowledgments

AS was supported by the European Commission under grant agreement no. 101135671. We thank the reviewers for their constructive and helpful comments, which have significantly improved the quality and clarity of our manuscript.

References

AI Sweden. 2023. AI Sweden's LLM AI Model License Agreement for GPT-SW3. Accessed: December 2023.

Babych, Bogdan and Anthony Hartley. 2003. Improving machine translation quality with automatic named entity recognition. In *Proceedings of the 7th International EAMT workshop on MT and other language technology tools, Improving MT through other language technology tools, Resource and tools for building MT at EACL 2003*, Columbus, Ohio.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bang, Yejin, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In Park, Jong C., Yuki Arase, Baotian Hu, Wei Lu, Derry Wijaya, Ayu Purwarianti, and Adila Alfa Krisnadhi, editors, *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali, November. Association for Computational Linguistics.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In Larochelle, H., M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Daems, Joke, Orphée De Clercq, and Lieve Macken. 2017. Translationese and post-editese: How comparable is comparable quality? *Linguistica Antverpiensia New Series-Themes in Translation Studies*, 16:89–103.

Freitag, Markus, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. Technical report, Google DeepMind. Technical Report.

Hendy, Amr, 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. *arXiv preprint arXiv:2302.09210*.

Hvidfeldt, Jón Brian. 2020. Faroese language to feature on Google Translate, 10. Accessed: December 2023.

Jiao, Wenxiang, Wenxuan Wang, Jen-tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is ChatGPT a Good Translator? Yes With GPT-4 As The Engine. *arXiv preprint arXiv:2301.08745*.

Lyu, Chenyang, Jitao Xu, and Longyue Wang. 2023. New trends in machine translation using large language models: Case examples with ChatGPT. *arXiv preprint arXiv:2305.01181*.

Meta AI. 2023. Llama 2 community license agreement. https://ai.meta.com/llama/license/. Accessed: December 2023.

Mistral AI team. 2023. Mixtral of experts: A high quality sparse mixture-of-experts. https://mistral.ai/news/mixtral-of-experts/. Accessed: January 2024.

OpenAI. 2023. Terms of use. https://openai.com/policies/terms-of-use. Accessed: December 2023.

Poncelas, Alberto, Dimitar Shterionov, Andy Way, Gideon Maillette de Buy Wenniger, and Peyman Passban. 2018. Investigating backtranslation in neural machine translation. In Pérez-Ortiz, Juan Antonio, Felipe Sánchez-Martínez, Miquel Esplà-Gomis, Maja Popović, Celia Rico, André Martins, Joachim Van den Bogaert, and Mikel L. Forcada, editors, *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, pages 269–278, Alicante, Spain, May.

Poncelas, Alberto, Maja Popović, Dimitar Shterionov, Gideon Maillette de Buy Wenniger, and Andy Way. 2019. Combining PBSMT and NMT back-translated data for efficient NMT. In Mitkov, Ruslan and Galia Angelova, editors, *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 922–931, Varna, Bulgaria, September. INCOMA Ltd.

Scalvini, Barbara and Iben Nyholm Debess. 2024. Evaluating the potential of language-family-specific generative models for low resource data augmentation: a Faroese case study. In *Proceedings of The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING '24)'*), Torino.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In Erk, Katrin and Noah A. Smith, editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany, August. Association for Computational Linguistics.

Símonarson, Haukur Barri, Vésteinn Snæbjarnarson, Pétur Orri Ragnarsson, Haukur Páll Jónsson, and Vilhjálmur Þorsteinsson. 2021. Miðeind's WMT 2021 submission. *arXiv preprint arXiv:2109.07343*.

Simonsen, Annika, Sandra Saxov Lamhauge, Iben Nyholm Debess, and Peter Juel Henrichsen. 2022. Creating a Basic Language Resource Kit for Faroese. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4637–4643.

Snæbjarnarson, Vésteinn, Annika Simonsen, Goran Glavaš, and Ivan Vulić. 2023. Transfer to a low-resource language via close relatives: The case study on Faroese. In Alumäe, Tanel and Mark Fishel, editors, *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 728–737, Tórshavn, Faroe Islands, May. University of Tartu Library.

Statistics Faroe Islands. 2024. Population. Accessed: February 2024.

Swedish Government. 2023. Regeringen tillsätter en AI-kommission för att stärka svensk konkurrenskraft, 12. Press release from Finansdepartementet, Statsrådsberedningen.

Talhadas, Paulo. 2023. Quality reports - monitor the quality results for your translations, 11. Accessed: December 203.

Team, NLLB, 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.

Vanmassenhove, Eva, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2203–2213.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Yang, Wayne and Garrett Nicolai. 2023. Neural Machine Translation Data Generation and Augmentation using ChatGPT. *arXiv preprint arXiv:2307.05779*.

Zhang, Mike and Antonio Toral. 2019. The effect of translationese in machine translation test sets. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 73–81, Florence, Italy, August. Association for Computational Linguistics.

A Quantitative Error Analysis

We count the most common errors and display them in Table 3. Furthermore, we refer to two additional categories of common errors found in GPT-4's translations from Faroese to English.

Translationese
GPT-4 occasionally generates translations that come across as awkward or grammatically incorrect in English, an issue commonly encountered in MT referred to as "machine-translationese" (Zhang and Toral, 2019; Daems et al., 2017; Vanmassenhove et al., 2021). In the case of the Faroese to English translations, GPT-4 sometimes opts for a literal, word-for-word translation approach, leading to syntax that sounds unnatural. For example:

- **FO:** Illgruni er tó um, at sjey onnur eisini eru smittað við nýggja frábrigdinum, **skrivar Ritzau.**
- **GPT4** "However, there is suspicion that seven others are also infected with the new variant, **writes Ritzau.**"

In this case, it is more natural to choose the word order, "Ritzau writes". However, it is worth to note that this type of error is possibly not thought of as an error by some, because it could in reality be a question of style-preference.

Inappropriate Register
Finally, GPT-4 sometimes opts for translations that carry an inappropriate tone, especially noticeable in formal settings such as news articles. For instance, *andaðist* is translated into the more colloquial "died" rather than the more fitting and respectful "passed away". This discrepancy in tone becomes particularly evident in news reporting, where a certain level of formality is anticipated. However, it is crucial to acknowledge that the register and genre of the text were not defined when prompting GPT-4.

Category	News Sentences (425)	Blog Sentences (425)
Correct-translation-wrong-terminology	89	20
NEs	26	5
Cultural context	16	18
Idioms and fixed phrases	11	19
DKK	19	0
Translationese	10	7
Icelandicism	6	7
Source error	8 (2)	2
Faroese	3	1
Inappropriate register	2	0
COVID	1	0
Other	41	56

Table 3: Detailed evaluation results for specific categories in translations of 425 news sentences and 425 blog sentences. The figures in parentheses indicate the count of translation errors within the total reported for that category.