Prompt Engineering Enhances Faroese MT, but Only Humans Can Tell

Barbara Scalvini
University of the Faroe Islands
barbaras@setur.fo

Annika Simonsen
University of Iceland
ans72@hi.is

Iben Nyholm Debess
University of the Faroe Islands
ibennd@setur.fo

Hafsteinn Einarsson
University of Iceland
hafsteinne@hi.is

Abstract

This study evaluates GPT-4's English-to-Faroese translation capabilities, comparing it with multilingual models on FLORES-200 and Sprotin datasets. We propose a prompt optimization strategy using Semantic Textual Similarity (STS) to improve translation quality. Human evaluation confirms the effectiveness of STS-based few-shot example selection, though automated metrics fail to capture these improvements. Our findings advance LLM applications for low-resource language translation while highlighting the need for better evaluation methods in this context.

1 Introduction

Historically, it has been a challenge to achieve high-quality machine translations (MT) for low-resource languages. The lack of resources has been shown to impact not only the development of high performing MT models, but also the development of high quality automated translation metrics (Callison-Burch et al., 2011; Bojar et al., 2014; Koehn and Knowles, 2017; Ranathunga et al., 2023). Low-resource languages often have to rely on string-based language independent metrics such as BLEU (Papineni et al., 2002) and ChrF (Popović, 2015). However, these methods have shown to perform poorly when compared to neural metrics, as shown by the WMT22 Metrics Shared Task (Freitag et al., 2022). The lack of neural metrics developed for these languages often leaves expensive and slow human evaluation as the only high quality alternative for detecting nuanced improvement in translation quality.

Recent advancements in LLMs offer opportunities to mitigate the effect that low-resources have on translation performance, leveraging few-shot learning to achieve remarkable performances with minimal data requirements (Brown et al., 2020). However, there is a disparity in the translation performance when it comes to low-resource languages vs high-resource languages (Hendy et al., 2023; Lyu et al., 2023; Bang et al., 2023; Chang et al., 2024). Therefore, optimizing the efficiency of these models in data-constrained environments demands a strategic approach in order to get the best performance. There is still much that is unknown about how prompt engineering and few-shot example selection influences translation performance. Furthermore, LLMs have proven to be a competitive alternative also for what concerns translation evaluation (Kocmi and Federmann, 2023). However, this ability of LLMs to assess translation has not been proven yet in the context of low-resource languages.

We investigate how STS-driven example selection, applied with the translation query, improves translation quality in GPT-4 Turbo (OpenAI et al., 2024), specifically the `gpt-4-1106-preview` release, for Faroese[1], a critically low-resource language. Our findings therefore demonstrate a novel approach to improve the utility of sparse data. Moreover, we demonstrate how current translation metrics cannot adequately capture nuances in translation performance, advocating for the development of more robust evaluation tools. Through this exploration, we provide an assessment of the state of the art of MT for Faroese, and highlight how current automated evaluation metrics cannot appropriately capture nuanced improvements provided by prompt engineering.

Our contributions are the following:

- Creating three synthetic parallel datasets with 1012 sentences each from the FLORES benchmark (NLLB Team et al., 2022),

[1]The population of the Faroe Islands is 54.000 (Statistics Faroe Islands, 2024).

Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025), pages 622–633
March 3-4, 2025

translated from English to Faroese by GPT-4 Turbo using zero-shot, random few-shot, and STS-based few-shot techniques respectively [2].

- Conducting an automated evaluation of these datasets employing BLEU, ChrF, and BERTScore metrics, alongside a GPT-4 Turbo confidence score for the few-shot datasets.
- Ranking of translations from each dataset on a subset of 200 sentences, performed by multiple native Faroese speakers and GPT-4. We further ranked 200 sentences sourced from another dataset to confirm our results.
- Benchmarking GPT-4 Turbo's English-Faroese translation performance against multilingual translation models covering Faroese such as MADLAD-400 and NLLB-200.

2 Previous Work

2.1 Machine Translation for Faroese

Historically, the limited amount of parallel data available for Faroese has hindered the development of MT tools for the language. However, in recent years, efforts have been made to address this issue and ensure better coverage of Faroese. One such effort led to the creation of the *Sprotin's parallel corpus* (Mikkelsen, 2021), a collection of around 100K English-Faroese human translated sentences. This corpus facilitated the inclusion of Faroese in *Microsoft Translator* and the development of a Faroese MT model, named *Vélþýðing* (Símonarson et al., 2021), by Miðeind, an Icelandic NLP company. The rise of massively multilingual translation models has sparked several initiatives aimed at including low-resource languages, thanks to their capability for cross-lingual transfer and the exploitation of shared linguistic features. Notably, initiatives such as Google's MADLAD 400 (Kudugunta et al., 2023) and Meta's No Language Left Behind (NLLB) (NLLB Team et al., 2022) target specifically low-resource languages, including Faroese. As of July 2024, Faroese is also included in Google Translate, Google's effort to develop an MT system for over 1,000 languages (Bapna et al., 2022). The development of these multilingual models still predominantly relies on string-based evaluation metrics like BLEU and ChrF. Despite the widespread criticism and the documented limitations of these metrics (Reiter, 2018; Callison-Burch et al., 2006) they continue to serve as the de facto standard in the field, particularly for low-resource languages, which are for the most part not included in shared tasks aimed at metrics evaluations (Freitag et al., 2022; Mathur et al., 2020). This persistence is likely due to their simplicity, ease of implementation, historical precedent, and, often, lack of affordable alternatives. The recent development of a BERT model for Faroese (Snæbjarnarson et al., 2023) has presented the opportunity to add BERTScore (Zhang et al., 2020), a metric based on contextual embeddings, to the pool of available metrics for Faroese.

2.2 The Rise of LLMs in Machine Translation

With the recent rise of LLMs it became apparent that transformer based MT models are not necessarily the go-to solution anymore when dealing with automatic translation. The few-shot learning capabilities of LLMs opened new avenues for translation with small data. Brown et al. (2020), with their paper titled "Language Models are Few-Shot Learners", demonstrated that GPT-3 could understand and execute tasks, including translation, with minimal examples through in-context learning (ICL). This capacity of LLMs to adapt to specific tasks with just a few guiding examples represents a shift in paradigm from traditional MT methods (Lyu et al., 2024), which often rely on extensive supervised training.

Recently, LLMs have revealed their potential not only as translator but also evaluators of translation (Karpinska and Iyyer, 2023; Fernandes et al., 2023; Huang et al., 2024), reaching state-of-the-art accuracy with respect to human evaluation (Kocmi and Federmann, 2023). However, these results were mostly obtained for high-resource languages, while the potential of LLMs for translating and evaluating translation of low-resource languages remains mostly untapped. In the specific case of Faroese, studies have already been conducted to assess how well LLMs understand the language within the context of MT. Scalvini and Debess (2024) evaluated the language comprehension capabilities of an LLM that targets Nordic

[2] https://huggingface.co/datasets/barbaroo/FLORES200_translations_GPT4

languages, GPT-SW3, while Debess et al. (2024) and Simonsen and Einarsson (2024) explored GPT-4's performance in Faroese sentiment analysis and translation from Faroese to English, where it showed good performance.

2.3 The Role of Semantic Textual Similarity in Prompt Engineering.

While some studies have focused on using a zero-shot prompting technique to translate, achieving performance comparable to those of conventional MT systems (Jiao et al., 2023; Chang et al., 2024), the potential of few-shot prompting, particularly in the realm of low-resource languages, invites further exploration. Prior research has predominantly relied on the use of randomly chosen translation examples as prompts. However, emerging studies have explored structured approaches, such as Pattern-Exploiting Training (Schick and Schütze, 2021), K-Nearest-Neighbour (kNN) selection for choosing translation examples from a pool of high-quality candidates (Vilar et al., 2022; Zhu et al., 2023) or choosing examples based on STS (Zhang et al., 2023). Such studies indicate that the quality of translation examples plays a crucial role in the effectiveness of LLMs for MT.

Despite these advancements, the effectiveness of using semantically similar translation examples in MT with LLMs remains an open question. Findings by Vilar et al. (2022) and Zhang et al. (2023) suggest that while example quality is crucial, STS alone does not strongly correlate with improved translation performance. On the other hand, other research, such as the study by Moslem et al. (2023) which utilizes lexical fuzzy matches to find similar translations, points towards significant benefits from employing semantically related examples. It is worth noting that most of this research has focused on high-resource language pairs and previous iterations of LLMs: these results might not therefore directly translate to current LLM versions and low-resource languages. Furthermore, most LLMs are capable of generating grammatically correct output in high-resourced languages, but often fail when zero-shot prompted in languages such as Faroese, making generative language tasks such as translation and summarization challenging. This discrepancy highlights the need for further investigation into the optimal use of example selection strategies in enhancing LLM-based translation into low-resource languages. Conditioning on grammatically correct and good translation examples has the potential to improve LLM generation quality for low-resourced languages.

3 Methods

3.1 Prompting GPT-4 for Translation

We prompted the GPT-4 Turbo model (`'gpt-4-1106-preview'`) (OpenAI et al., 2024) for English to Faroese translation in a zero and few-shot setting. This model was selected based on its superior performance in Faroese language generation at the time, as evidenced by preliminary experiments made by the authors of this paper. The prompting strategies used are described below:

- Zero-shot setting.
- Few-shot setting with random selection of 12 parallel sentences from the *Sprotin* dataset (Mikkelsen, 2021). We will refer to such translations as T_{rand}.
- Few-shot setting with selection of 12 parallel sentences from the *Sprotin* dataset based on the highest STS with the translation query (T_{sel}). Note that the translation query is in English, so the similarity search is based on English examples. Their Faroese translated sentences are then used in the few-shot prompt.

The Sprotin dataset is, to our knowledge, the largest collection of high quality human translated English-Faroese sentences pairs. STS was quantified by a multilingual model, `Multilingual-E5-large` (Wang et al., 2024), which was the highest ranking multilingual embedding model at the time according to the MTEB leader board[3]. The system prompt specified the proficiency of the chat-bot in the Faroese language ('*You are an expert in the Faroese language*') and the desired translation quality ('*The translations should be of excellent quality*'). With these settings, we translated from English to Faroese the test split of the FLORES dataset (NLLB Team et al., 2022), comprised of 1012 sentences.

[3] `https://huggingface.co/spaces/mteb/leaderboard`

3.2 Comparison with SOTA MT Models

We benchmark GPT-4's performance against two state of the art multilingual MT models, MADLAD-400 (10B parameters) and NLLB-200 (3B parameters). At the time of writing this paper, the Google Translate API did not allow access to the latest model, covering Faroese. Google Translate was therefore not included in the analysis. The models were used out of the box, without any fine-tuning for the English-Faroese pair, to translate the test split of the FLORES-200 dataset.

3.3 Evaluation on the FLORES-200 Test Set

In order to evaluate and compare translations, several metrics were used: two string-based metrics, BLEU (Papineni et al., 2002) and ChrF score (Popović, 2015), and one neural metric, BERTScore (Zhang et al., 2023), and a human evaluation score. Additionally, GPT-4 was asked to provide a score estimating how confident it was in the translation produced. The BERT model provided to BERTScore for evaluation was, to the best of our knowledge, the only available BERT model specifically catering to Faroese, FoBERT (Snæbjarnarson et al., 2023). We did not find any other neural metric that includes Faroese among its target languages, a situation common to most low and critically low-resource languages. Human evaluation was carried out by three linguists who are native speakers of Faroese. These experts ranked the four Faroese translations - the human translation from FLORES, the zero shot translations, T_{rand} and T_{sel} - blindly from best to worst (1 to 4) (see Figure 1 for an example of the annotation setup in Google Sheets). Annotators were presented with an error type hierarchy to align ranking criteria. According to the hierarchy, sentences with major errors like incomplete translations or lexical errors will be ranked lower than sentences with minor errors such as incorrect inflection or spelling errors. The human evaluation was performed on a subset of 200 translation queries, randomly selected. In this subset, 12 sentences were found which yielded two or more identical translations obtained by different translation methods (zero-shot, T_{rand}, T_{sel} or human reference). These were given the same rank by the annotators [4]. The annotators evaluated the same examples, so that inter-annotator agreement could be compared. For comparison, we asked GPT-4 to perform the same ranking task, over the same subset of sentences.

[4]the ranking could then be 1, 1, 3, 4 if the top ranked sentences are identical or 1, 2, 2, 4 if the second place translations are identical and lastly 1, 2, 3, 3 if the last rankings are identical

3.4 Replication on Another Source

In order to test the robustness of our human evaluation procedure and its findings, we selected 200 sentences randomly from the Sprotin corpus, and translated them following the three translation strategies presented in Section 3.1, with the aim to reproduce human ranking on this subset. However, the nature of the Sprotin sentences led us to reconsider our strategy: Sprotin is for the most part composed by short, simple, everyday sentences. Such sentences ended up being translated identically across translation strategies, leading to 132 sentences out of 200 having at least two identical translations, 39 having 3 identical sentences, and 21 having all 4 identical entries (three GPT-4 translation strategies plus the human translation). We considered the ranking of these entries a challenging - if not impossible - task, and therefore decided to change selection strategy for test sentences. Preliminary results from this evaluation attempt are discussed in Section 4. Subsequently, we decided to select 200 sentences randomly among longer sentences, as defined by number of tokens in the translation query. The threshold for selection was 18 tokens, as identified by rounding up the average number of tokens in a Sprotin sentence (8.8 tokens) plus 2 standard deviations (8.5). The rationale behind this choice is that longer sentences are more likely to be more linguistically complex and present more opportunities for variation in translation quality. The final subset presented an average of 28.6 tokens, roughly 2 tokens longer than that of FLORES (26.8). This selection thus brought the two subset closer together in terms of average sentence length. These 200 sentences were then translated according to the three different translation strategies, and translations were ranked by two human evaluators from best to worst (1 to 4). Out of 200 translation queries, 4 sentences were not parsed correctly by GPT-4, yielding to incomplete translations. These sentences were excluded from the analysis.

3.5 Annotator Agreement

To assess the degree of agreement among the raters for the ranking tasks, we employed Kendall's Coefficient of Concordance (W). This

non-parametric statistic is particularly suited for situations where three or more raters are asked to order a set of items, as it measures the extent of agreement among the raters' rankings (Kendall, 1938). Kendall's W ranges from 0, indicating no agreement, to 1, denoting perfect concordance.

Each assignment consisted of three Faroese native speakers for FLORES and two native speakers for Sprotin providing rankings for four items. For each assignment, we calculated Kendall's W to determine the level of rater agreement. We then computed the average of these values across all assignments to obtain an overall measure of agreement. This approach allowed us to quantify the consistency of raters' evaluations across multiple independent tasks, providing a robust assessment of inter-rater reliability in the context of our study.

4 Results

4.1 Automated Metrics are Blind

Automated metrics such as ChrF, BLEU and BERTScore reveal that GPT-4 produces translations of higher quality with respect to the two MT models, MADLAD-400 and NLLB-200 (see Table 1) on the FLORES dataset. However, when it comes to comparing the different GPT-4 prompting strategies in terms of translation performance, these metrics appear to be "blind" to subtle improvements. By "blind," we mean that the automated metrics are not picking up on the improvement in performance when using the selected method (T_{sel}) over random (T_{rand}) - an improvement that is evident to human evaluators. Statistical comparison between the ChrF, BLEU and BERTScore distributions revealed no statistical difference in translation quality between zero-shot translation, T_{rand} and T_{sel}.

4.2 Human Evaluation on FLORES

Human evaluation revealed a small, but statistically significant difference between T_{rand} and T_{sel}. As Figure 2 shows, human evaluation was consistently ranked first, followed by the STS driven few-shot translation. We aggregated the rankings of T_{rand} and T_{sel} produced by the three evaluators and compared them statistically by Mann–Whitney U test, yielding a p-value of 0.006, indicating that the two distributions are indeed dissimilar. Interestingly, a slightly less substantial difference was found comparing T_{rand} with zero-shot translations ($p = 0.026$), indicating that providing random examples is a useful approach, but there is still a margin of improvement in translation quality to be exploited by example selection and prompt optimization. To summarize, Table 2 reports how many times (in percentage) each approach received the highest rank. Although the human translation was found to be superior the vast majority of instances, we see that each approach was occasionally ranked first, indicating how nuanced the differences between the approaches can be, when in a low-resource scenario.

4.3 Replication on the Sprotin Subset

As mentioned in Section 3, translating randomly selected sentences from the Sprotin Corpus resulted in many identical translation entries, rendering the set unsuitable for ranking. However, to gain preliminary insight into the performance of the different translation methods on this subset, we counted how often each strategy produced output identical to the human translation. Interestingly, we found that T_{sel} produced the highest number of human-like translations (47), followed by T_{rand} (36) and zero-shot (31). When considering the human reference as the gold standard, these preliminary results mirror the hierarchy observed in the human ranking of the FLORES sentences. The second round of evaluation, concerning the ranking of longer sentences extracted randomly from Sprotin, showed compatible results with our previous findings over the FLORES dataset (Figure 3). We see the human entry being consistently ranked first, obtaining an overall ranking of 1.5, and the zero-shot approach being ranked last overall. The difference between T_{rand} and T_{sel} is however more pronounced in the Sprotin subset than what we observed in FLORES. Statistical comparison of the two distributions yields a p-value of 1.27 e^{-6}, a strong evidence that the two distributions are in fact distinct, and that the T_{sel} strategy produces statistically better translations. If we take a look at Table 1, displaying how many times each approach was ranked first, in percentage, we find remarkably consistent results between FLORES and Sprotin, a result which supports the robustness of our method and findings.

4.4 Annotator Agreement

The average Kendall's W value obtained was 0.694 for FLORES and 0.752 for Sprotin, indicating a substantial level of agreement among the raters, which supports the reliability of the ranking

English Sentence	Translation	Evaluation
['Before The Simpsons Simon had worked on several shows in various positions.']	Áðrenn The Simpsons arbeiddi Simon við fleiri sýningum í ymiskum støvum.	3
['Before The Simpsons Simon had worked on several shows in various positions.']	Áðrenn The Simpsons starvaðist Simon á ymiskum sjónvarpssendingum í ymsum støvum.	1
['Before The Simpsons Simon had worked on several shows in various positions.']	Áðrenn The Simpsons høvdu Simon arbeitt við ymiskum sendingum í ymsum støvum.	4
['Before The Simpsons Simon had worked on several shows in various positions.']	Áðrenn The Simpsons hevði Simon arbeitt við fleiri sendingar í ymiskum starvum.	2

Figure 1: Example of human evaluation setup in a spreadsheet where 4 is the lowest and 1 is the highest rank.

Translation Method	BLEU	ChrF	BERTScore F1
MADLAD-400	13.62 ± 0.53	40.89 ± 0.54	$0.9373 \pm 8 \times 10^{-4}$
NLLB-200	16.79 ± 0.52	48.05 ± 0.39	$0.9474 \pm 5 \times 10^{-4}$
Zero-shot GPT-4	21.36 ± 0.50	52.55 ± 0.39	$0.9516 \pm 5 \times 10^{-4}$
T_{rand} few-shot GPT-4	21.09 ± 0.49	52.36 ± 0.38	$0.9515 \pm 5 \times 10^{-4}$
T_{sel} few-shot GPT-4	21.77 ± 0.50	53.24 ± 0.38	$0.9524 \pm 5 \times 10^{-4}$

Table 1: Translation performance of MADLAD-400, NLLB-200, and GPT-4 on the FLORES-200 dataset for English to Faroese translations.

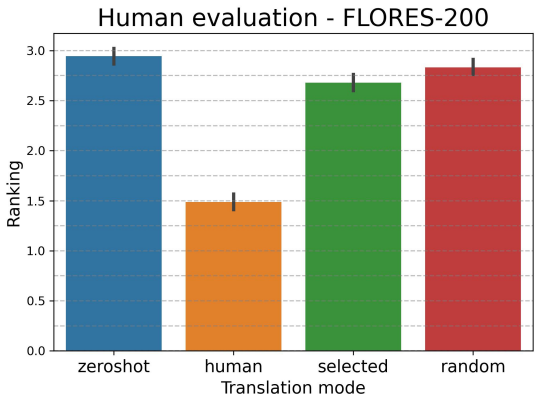

Figure 2: Human evaluation results, for a subset of 200 FLORES sentences. Translations were ranked from best to worst (1 to 4). The T_{rand} (random) and T_{sel} (selected) distributions are statistically different, yielding a p-value of 0.006 by Mann-Whitney U test.

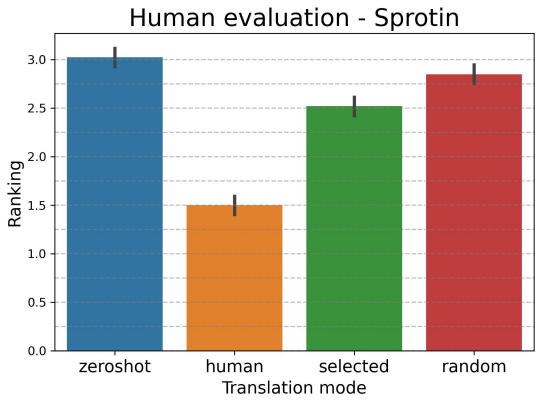

Figure 3: Human evaluation results, for a subset of 200 Sprotin sentences. Translations were ranked from best to worst (1 to 4). The T_{rand} (random) and T_{sel} (selected) distributions are statistically different, yielding a p-value of 1.27 e^{-6} by Mann-Whitney U test.

data used in our analyses.

4.5 GPT-4 is Also Blind

The confidence score provided by GPT-4 was in alignment with human evaluation for what concerns the presence of a statistical difference between T_{rand} and T_{sel} (p value = 1.7 e^{-10}), as can be seen in Figure 5. It is however important to notice how GPT-4 output a confidence score of 0.95 for 93% per cent of translations, a result which is in line with previous findings by Kocmi and Federmann (2023). While these results align with human evaluation, the characteristics of such a distribution make comparison by statistical analysis less reliable.

To further investigate GPT-4' s understanding of translation nuances, we prompted GPT-4 for translation ranking in a setting that mimics that of human evaluation: the chatbot was asked to rank the 4 translation option from best to worst (1 to 4), on the same set of translated sentences evaluated by human experts. Notably, GPT-4 fails to identify human translation as the best one (Figure 4). Specifically, GPT-4 ranked T_{rand} statistically higher than T_{sel} (p value = 0.026) and human translation (p value = 7.3e-4). This result there-

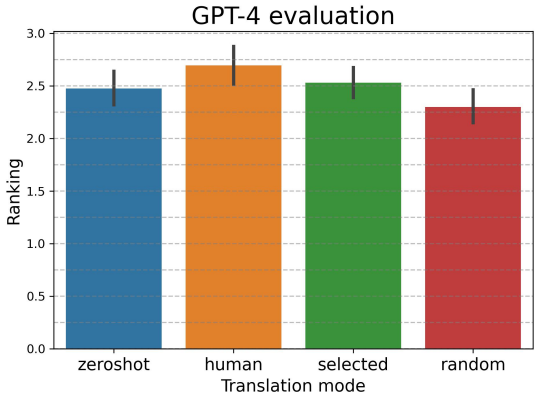

Figure 4: Evaluation rankings assigned by GPT-4, for a subset of 200 FLORES sentences. Translations were ranked from best to worst (1 to 4).

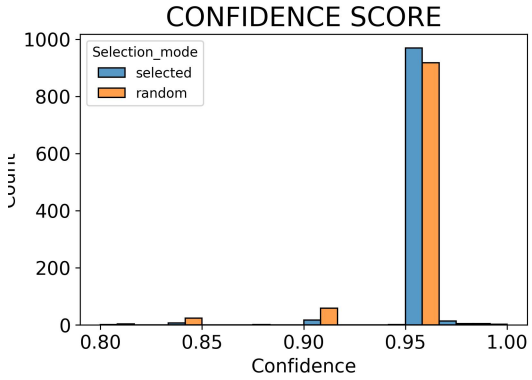

Figure 5: Confidence score assigned by GPT-4 to its own translations of the devtest split of the FLORES-200 dataset. Values for the two distributions are plotted side by side for ease of visualization. The labels 'selected' and 'random' refer respectively to T_{sel} and T_{rand} few-shot translations.

fore shows that GPT-4 is also blind to subtle improvements in translation quality and once again underlines how automated metrics degrade in performance in a low-resource setting.

5 Discussion

5.1 Challenges in Evaluating Low-Resource Language Translation

During our study, we observed how prompt engineering can in fact provide improvements in translation quality into low-resource languages such as Faroese. In order to prove this, we used STS-based few-shot prompting as a proof of concept. While human evaluators were able to detect such improvement, automated scores available for the language, BLEU, ChrF and BERTScore, failed to do so. That being said, among the automated metrics used, BLEU was most sensitive in detecting the improvement of the selected method (T_{sel}) over the random method (T_{rand}), albeit the difference was small (see Table 1), with overlapping confidence intervals, indicating that it was not able to tell if there was an improvement. In addition to utilizing the above mentioned automated metrics and human evaluation, we also utilize a GPT-4 based confidence score, which is a way to evaluate translation performance from the model's own perspective. We hypothesize that prompt engineering driven improvements are too nuanced to be detected by currently available automated metrics, including string-based metrics (BLEU, ChrF) and BERTScore. GPT-4's evaluation also presented critical pitfalls, showing how the model prefers its own output with respect to the human reference. Higher performance automated metrics such as COMET and UNITE (Freitag et al., 2022) are not available for Faroese and for the majority of low-resource languages, as these neural-based metrics require specific resources like large, high-quality datasets for their development. Translation into Faroese and related quality evaluation poses multiple challenges, as Faroese is not only low-resource but also a morphologically rich language. Evaluating MT for morphologically rich languages is notoriously difficult due to the complexity and variability in word forms. These difficulties are well-documented in the literature, with studies highlighting the shortcomings of traditional evaluation metrics when applied to such languages (Freitag et al., 2022). While it is true that LLMs provide new opportunities for low-resource languages, such opportunities cannot be fully taken advantage of for a lack of appropriate methods to assess related improvements. In alignment with statements from Chang et al. (2024) and Sai et al. (2020), our findings highlight how automated metrics do not capture the nuances in quality as human evaluators do. Therefore, we strongly advocate for the development of more robust evaluation tools tailored to low-resource contexts, and in general, for the extension of neural metrics to low-resource languages.

Translation Method	Zero-shot	T_{rand} few-shot	T_{sel} few-shot	Human translation
FLORES - First-Rank (%)	7.83	7.33	11.67	74.33
SPROTIN - First-Rank (%)	7.14	7.65	12.75	74.23

Table 2: Percentage of times the four different translation strategies (human, zero-shot, T_{rand} and T_{sel}) were ranked first during human evaluation. Rankings for all evaluators were aggregated in the final percentage.

5.2 Significance of Semantic Textual Similarity in Few-shot

Our results demonstrate a small yet statistically significant improvement in GPT-4's translation quality of English to Faroese when using semantically similar examples, as highlighted by human evaluation. This improvement underscores GPT-4's ability to utilize the context that is provided by semantically similar examples to generate better translations. By using semantically similar examples effectively, our study demonstrates a potential pathway to achieve higher-quality translations without the need for an overly large dataset. Furthermore, We observed a stronger impact of example selection in the Sprotin subset, with respect to FLORES. This might be due to several factors. One possible aspect to consider is the type of language and domains found in FLORES, which are sometimes technical and not representative of every day speech. Therefore, the Sprotin sentences might present a better match to the examples (as they are extracted from the same dataset). Moreover, FLORES is a well known, widely available test dataset for translation, and there is a non negligible possibility of it being already included in GPT-4's training data. Had the model seen FLORES already, that would limit the impact of the prompting strategy on translation quality. Our findings also contribute to the broader understanding of prompt engineering, specifically in the context of low-resource languages. There is a benefit to selecting STS-based examples. Findings from previous work about the impact of STS are ambiguous (Vilar et al., 2022; Zhang et al., 2023; Moslem et al., 2023). However, they were mostly carried out on high resource languages, for which GPT-4's performance is generally of high quality. Therefore, we could reasonably expect a smaller margin of improvement, which is harder to detect unambiguously.

5.3 Limitations and Future Works

Our study, while insightful, has certain limitations that pave the way for future research. The focus on a single LLM and language constrains generalizability. Moreover, human evaluation introduces potential biases, particularly in identifying human-written translations. The datasets used lack Faroese cultural elements, and we cannot rule out the possibility of GPT-4 Turbo having been trained on the FLORES dataset. To address these limitations and expand our understanding, future work should explore multiple LLMs, including smaller and domain-specific models, and extend to other low-resource languages. This broader approach could improve the evaluation process and provide insights into the relationship between translation quality and corpus characteristics. Experimenting with an increased number of semantically similar examples and longer paragraphs for translation could enhance quality and offer a more comprehensive evaluation. As open-source models for low-resource languages improve, comparing their performance using our semantic similarity approach could be valuable. Lastly, studying the impact of reference corpus size and domain specificity on STS performance could deepen our understanding in diverse linguistic contexts.

6 Conclusion

This study shows that selecting few-shot learning examples based on STS can improve GPT-4 Turbo's Faroese translation performance, as confirmed by human evaluation. However, current automated metrics fail to detect these improvements, highlighting a critical issue in low-resource language translation evaluation. While LLMs offer new opportunities for language generation, the inability of automated metrics to capture progress in low-resource contexts could widen digital language representation disparities. This situation necessitates expensive human evaluation, potentially hindering advancements. Therefore, we call

for collaborative efforts to develop metrics specifically designed for low-resource language contexts.

References

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. *arXiv preprint arXiv:2302.04023*.

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Baljekar, Xavier Garcia, Wolfgang Macherey, Theresa Breiner, Vera Axelrod, Jason Riesa, Yuan Cao, Mia Xu Chen, Klaus Macherey, Maxim Krikun, Pidong Wang, Alexander Gutkin, Apurva Shah, Yanping Huang, Zhifeng Chen, Yonghui Wu, and Macduff Hughes. 2022. Building Machine Translation Systems for the Next Thousand Languages.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, et al. 2014. Findings of the 2014 Workshop on Statistical Machine Translation. In *Proceedings of the ninth workshop on statistical machine translation*, pages 12–58.

Tom Brown, 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 *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar F. Zaidan. 2011. Findings of the 2011 Workshop on Statistical Machine Translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, WMT '11, page 22–64, USA. Association for Computational Linguistics.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the Role of Bleu in Machine Translation Research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256, Trento, Italy. Association for Computational Linguistics.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A Survey on Evaluation of Large Language Models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Iben Nyholm Debess, Annika Simonsen, and Hafsteinn Einarsson. 2024. Good or Bad News? Exploring GPT-4 for Sentiment Analysis for Faroese on a Public News Corpora. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7814–7824, Torino, Italia. ELRA and ICCL.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André Martins, Graham Neubig, Ankush Garg, Jonathan Clark, Markus Freitag, and Orhan Firat. 2023. The Devil Is in the Errors: Leveraging Large Language Models for Fine-grained Machine Translation Evaluation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1066–1083, Singapore. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chikiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 Metrics Shared Task: Stop Using BLEU – Neural Metrics Are Better and More Robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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.

Xu Huang, Zhirui Zhang, Xiang Geng, Yichao Du, Jiajun Chen, and Shujian Huang. 2024. Lost in the Source Language: How Large Language Models Evaluate the Quality of Machine Translation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 3546–3562, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Wenxiang Jiao, 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*.

Marzena Karpinska and Mohit Iyyer. 2023. Large Language Models Effectively Leverage Document-level Context for Literary Translation, but Critical Errors Persist. In *Proceedings of the Eighth Conference on Machine Translation*, pages 419–451, Singapore. Association for Computational Linguistics.

M. G. Kendall. 1938. A New Measure of Rank Correlation. *Biometrika*, 30(1/2):81–93.

Tom Kocmi and Christian Federmann. 2023. Large Language Models Are State-of-the-Art Evaluators

of Translation Quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Philipp Koehn and Rebecca Knowles. 2017. Six Challenges for Neural Machine Translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39. Association for Computational Linguistics. First Workshop on Neural Machine Translation, NMT 2017 ; Conference date: 04-08-2017 Through 04-08-2017.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A. Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. MADLAD-400: A Multilingual And Document-Level Large Audited Dataset.

Chenyang Lyu, Zefeng Du, Jitao Xu, Yitao Duan, Minghao Wu, Teresa Lynn, Alham Fikri Aji, Derek F. Wong, and Longyue Wang. 2024. A Paradigm Shift: The Future of Machine Translation Lies with Large Language Models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1339–1352, Torino, Italia. ELRA and ICCL.

Chenyang Lyu, Jitao Xu, and Longyue Wang. 2023. New Trends in Machine Translation using Large Language Models: Case Examples with ChatGPT. *arXiv preprint arXiv:2305.01181*.

Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020. Results of the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725, Online. Association for Computational Linguistics.

Jonhard Mikkelsen. 2021. Sprotin sentences. `https://raw.githubusercontent.com/Sprotin/translations/main/sentences_en-fo.strict.csv`. Accessed: October 13, 2023.

Yasmin Moslem, Rejwanul Haque, John D Kelleher, and Andy Way. 2023. Adaptive Machine Translation with Large Language Models. *arXiv preprint arXiv:2301.13294*.

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.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov,

Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. GPT-4 Technical Report.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural Machine Translation for Low-resource Languages: A Survey. *ACM Comput. Surv.*, 55(11).

Ehud Reiter. 2018. A Structured Review of the Validity of BLEU. *Computational Linguistics*, 44(3):393–401.

Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. 2020. A Survey of Evaluation Metrics Used for NLG Systems. ArXiv:2008.12009 [cs].

Barbara Scalvini 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 2024)*, pages 6496–6503, Torino, Italia. ELRA and ICCL.

Timo Schick and Hinrich Schütze. 2021. Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Haukur Barri Símonarson, Vésteinn Snæbjarnarson, Pétur Orri Ragnarson, Haukur Jónsson, and Vilhjalmur Thorsteinsson. 2021. Miðeind's WMT 2021 submission. In *Proceedings of the Sixth Conference on Machine Translation*, pages 136–139, Online. Association for Computational Linguistics.

Annika Simonsen and Hafsteinn Einarsson. 2024. A Human Perspective on GPT-4 Translations: Analysing Faroese to English News and Blog Text Translations. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation*, volume 1, pages 24–36, Sheffield, United Kingdom. Research And Implementations & Case Studies.

Vésteinn Snæbjarnarson, Annika Simonsen, Goran Glavaš, and Ivan Vulić. 2023. Transfer to a Low-Resource Language via Close Relatives: The Case Study on Faroese. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 728–737, Tórshavn, Faroe Islands. University of Tartu Library.

Statistics Faroe Islands. 2024. Population. `https://hagstova.fo/en/population/population/population`. Accessed: May 2024.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2022. Prompting paLM for Translation: Assessing Strategies and Performance. *arXiv preprint arXiv:2211.09102*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual E5 Text Embeddings: A Technical Report.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting Large Language Model for Machine Translation: A Case Study. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 41092–41110. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis. *arXiv preprint arXiv:2304.04675*.